# Analyse classificatoire d'une correspondance multiple ; typologie et regression

Israël-César Lerman

IRIA

CENTRE DE RENNES

IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. 954 90 20

Rapports de Recherche

N° 194

# ANALYSE CLASSIFICATOIRE D'UNE CORRESPONDANCE MULTIPLE; TYPOLOGIE ET REGRESSION

Israël César LERMAN

Février 1983

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE
Tél. : (99) 36.20.00
Télex : UNIRISA 95 0473 F

# ANALYSE CLASSIFICATOIRE D'UNE CORRESPONDANCE MULTIPLE ; TYPOLOGIE ET REGRESSION

I.C. LERMAN

Publication Interne n° 186 - Janvier 1983
54 pages

RESUME : Nous avons ces dernières années développé, avec notamment la collaboration de B. TALLUR, la classification hiérarchique des lignes ou colonnes d'une juxtaposition de tables de contingence.

Le point de vue reste conforme à notre démarche générale dans l'évaluation des proximités entre structures statistiques. Cette approche s'est avérée particulièrement féconde aussi bien au niveau des résultats concrets obtenus qu'à celui des apports méthodologiques issus de cette structure particulière de la donnée :

- Définition de nouveaux indices d'association entre lignes ou classes de lignes (resp. colonnes ou classes de colonnes) qui se réfèrent soit à une représentation euclidienne, soit à une représentation ensembliste des éléments à comparer.

- Analyse statistique de croisements d'un type nouveau mettant en correspondance des classes de modalités, dont chacune est prise globalement.

- Régression qualitative entre une suite ordonnée de profils d'attitude -dont chacun est défini par une classe d'attributs s- et la suite des modalités d'une variable qualitative ordinale.

SUMMARY : We have, in the past few years, developped with B. TALLUR (and others), the hierarchical classification of rows or columns of a juxtaposition of contingency tables.

This research has been done according to our method for measuring proximities between statistical structures. The approach has been particulary fruitful with respect two aspects : the concrete results obtained on numerous analysis of real data and methodological contributions issued from the particular structure of the data.

Relative to the second aspect, we have obtained

- New elaboration of association's coefficients between rows or disjoint classes of rows (resp. columns or disjoint classes of columns). These coefficients take into account an euclidian or set theoretic representation of the statistical structures to be compared.

- Statistical analysis of new type of crossing fuzzy discreete variables putting in correspondence disjoint classes of attributes.

- Qualitative regression between an ordered sequence of behaviour's profiles -each one being defined by a class of attributes- and the ordered sequence of the modalities of an ordinal qualitative variable.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

# ANALYSE CLASSIFICATOIRE D'UNE CORRESPONDANCE MULTIPLE ; TYPOLOGIE ET REGRESSION

## I. RAPPEL DE LA CONSTRUCTION D'UN INDICE D'ASSOCIATION ENTRE VARIABLES QUALITATIVES.

Nous avons besoin pour commencer de rappeler notre démarche dans l'élaboration d'un indice d'association (nous disons encore de proximité, de similarité ou de ressemblance) entre variables qualitatives de même type. Nous le ferons pour deux types distincts qui interviendront nécessairement dans le cœur du sujet traité.

### I.1. Comparaison de deux attributs.

Désignons par $I$ l'ensemble des individus ou objets et par $(a, b)$ le couple d'attributs descriptifs à comparer, par $(I(a), I(b))$ le couple de parties de $I$ où $I(a)$ (resp. $I(b)$) est l'ensemble des individus qui possèdent l'attribut $a$ (resp. $b$).

Relativement à notre point de vue de représentation ensembliste des variables, les attributs $a$ et $b$ sont figurés par deux points de l'ensemble $\mathcal{P}(I)$ des

parties de $I$. La situation peut être naïvement schématisée comme suit

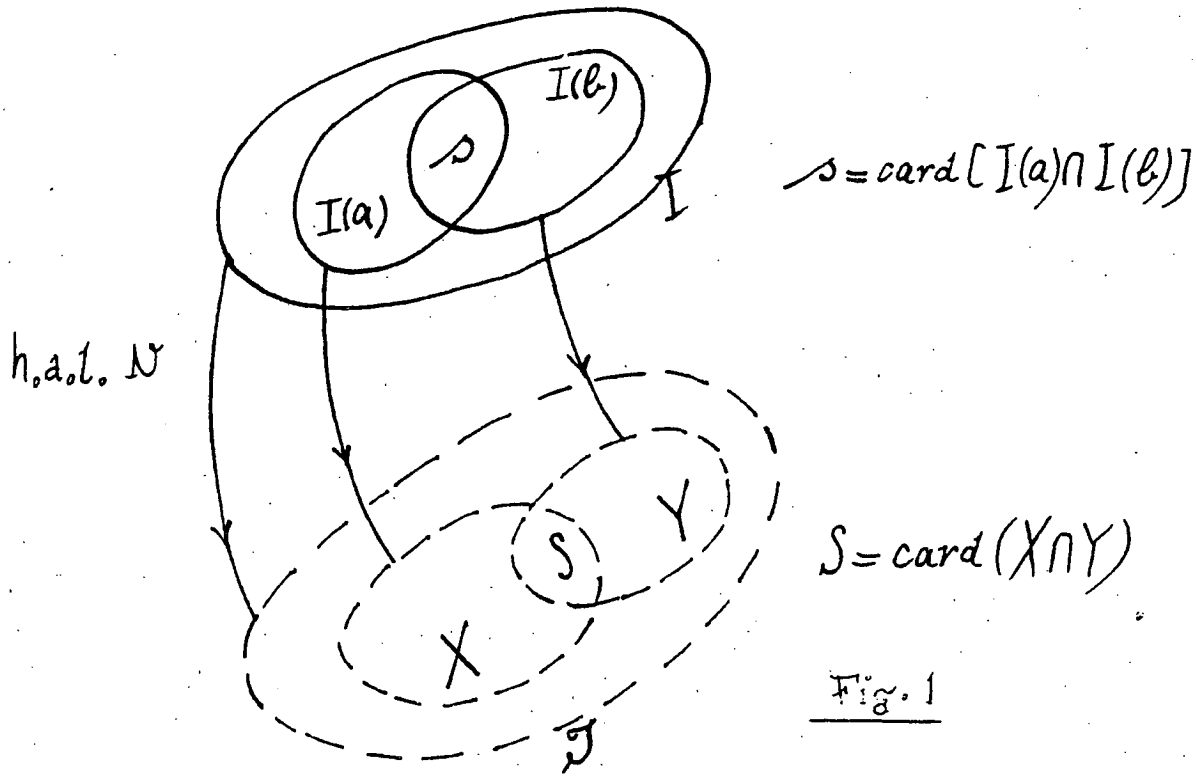



Fig. 1

Après avoir introduit l'indice brut $s = \text{card}[I(a) \cap I(b)]$, on considère une h.a.l.[(r)] $N$ qui associe au triplet $\{I ; I(a), I(b) / I(a) \subset I, I(b) \subset I\}$, un triplet d'ensembles aléatoires $\{J ; X, Y / X \subset J, Y \subset J\}$.

L'h.a.l. doit d'une « certaine façon » respecter les caractéristiques cardinales de $I(a)$, $I(b)$ et $I$ et à cet égard, nous avons pu dégager trois formes fondamentales de l'h.a.l. $N$ : $N_1$, $N_2$ et $N_3$ [LERMAN (1981a) chap. 2].

Pour $N_1$, $J = I$ et $X$ (resp. $Y$) est un élément aléatoire dans l'ensemble $\mathcal{P}_{n(a)}(I)$ (resp. $\mathcal{P}_{n(b)}(I)$), muni d'une probabilité uniformément répartie, des parties de $I$ de même cardinal $n(a) = \text{card}(I(a))$ (resp. $n(b) = \text{card}(I(b))$).

(r) hypothèse d'absence de lien.

D'autre part, $X$ et $Y$ sont indépendants. Dans ces conditions la v.a. $S = \mathrm{card}(X \cap Y)$ est hypergéométrique de moyenne $n(a)n(b)/n$ et de variance $n(a)n(\bar{a})n(b)n(\bar{b})/n^2(n-1)$. L'indice centré réduit est, au coefficient $\sqrt{n-1}$ près, l'indice d'association de K. Pearson.

Pour $N_2$, $\mathcal{I} = I$. Le choix de $X$ (resp. $Y$) se fait selon un modèle aléatoire à deux pas :

— le premier consiste dans le choix d'un niveau $k$ (resp. $h$) de $\mathcal{P}(I)$ avec une probabilité binomiale

$$\binom{n}{k} \alpha^k (1-\alpha)^{n-k}, \text{ où } \alpha = n(a)/n \text{ et } 0 \leqslant k \leqslant n,$$

(resp. $\binom{n}{h} \beta^h (1-\beta)^{n-h}$, où $\beta = n(b)/n$ et $0 \leqslant h \leqslant n$.). (1)

— le deuxième pas consiste dans le choix uniformément au hasard d'un élément qui est un $k$- sous ensemble (resp. $h$- sous ensemble) de $I$, à ce niveau.

De plus $X$ et $Y$ sont indépendants. Dans ces conditions, on démontre que $S = \mathrm{card}(X \cap Y)$ est une v.a. binomiale de paramètre $\pi = \alpha\beta$. Sa moyenne est donc la même que dans le cas de l'h.a.l. $N_1$, mais sa variance est égale à $n\pi(1-\pi)$.

Pour $N_3$, le choix de $X$ (resp. $Y$) se fait selon un modèle aléatoire à trois pas :

— le premier consiste à associer à $I$ un ensemble aléatoire $\mathcal{I}$, mais où l'aléa ne concerne que la cardinalité de $\mathcal{I}$. On suppose que $\nu = \mathrm{card}(\mathcal{I})$ est une v.a. de Poisson de paramètre $n = \mathrm{card}(I)$ :

$$\Pr\{\nu = \ell\} = \frac{n^\ell}{\ell!} e^{-n} \qquad (2)$$

— pour $\mathcal{U} = E_0$ fixé, les deux autres pas sont analogues.

à ceux de $N_2$. Les parties aléatoires X et Y étant indépendantes, on démontre que la v.a. $S = card(X \cap Y)$ suit une loi de Poisson de paramètre $n(a)n(b)/n$. On voit que la moyenne de S est la même que dans les deux cas précédents ; mais la variance devient ici égale à $n(a)n(b)/n$.

I.2 - Comparaison de deux variables qualitatives ordinales.

Ce procédé de construction d'un indice d'association entre attributs se généralise de façon naturelle à la comparaison de deux variables qualitatives de toutes sortes [LERMAN (1981a) Chap. 2]. Il se généralise également à la définition de coefficients d'association partielle pour les différents types de variables qualitatives [LERMAN (1983a) et (1983b)].

Nous allons - encore une fois parce que nous en aurons ci-dessous besoin et d'ailleurs, avec un codage mathématique adapté - reprendre l'introduction et les résultats de la comparaison de deux variables qualitatives ordinales.

$\{ C_\ell / 1 \leqslant \ell \leqslant L \}$ (resp. $\{ D_m / 1 \leqslant m \leqslant M \}$) est la suite des classes définie par l'une (resp. l'autre) variable qualitative ordinale. Les préordres totaux respectivement associés sont notés $\omega$ et $\varpi$. On représente $\omega$ et $\varpi$ par les sous ensembles suivants de $I \times I$ :

$$R(\omega) = \bigsqcup \{ C_\ell \times C_{\ell'} / 1 \leqslant \ell < \ell' \leqslant L \}$$

et

$$R(\varpi) = \bigsqcup \{ D_m \times D_{m'} / 1 \leqslant m < m' \leqslant M \} . \qquad (3)$$

Jusqu'à présent, nous avons directement travaillé avec les f.i. (fonctions indicatrices) de $R(\omega)$ et de $R(\varpi)$ que nous notons respectivement

$$\{ F_{ij} / (i,j) \in I^{[2]} \} \text{ et } \{ n_{ij} / (i,j) \in I^{[2]} \} , \qquad (4)$$

où $I^{[2]} = \{ (i,i') / i, i' \in I , i \neq i' \}$.

Nous serons ici amenés (cf. § III.3) à utiliser et étendre les f.i. des classes $C_\ell$ $(1 \leqslant \ell \leqslant L)$ et $D_m$ $(1 \leqslant m \leqslant M)$.

Si $\varphi_{C(\ell)}$ désigne la f.i. de $C_\ell$, $1 \leqslant \ell \leqslant L$ ; on a

$$(\forall (i,j) \in I^{[2]}), \quad \xi_{ij} = \sum \{ \varphi_{C(\ell)}^{(i)} \varphi_{C(\ell')}^{(j)} / 1 \leqslant \ell < \ell' \leqslant L \}. \quad (5)$$

De même, si $\varphi_{D(m)}$ désigne la f.i. de $D_m$, $1 \leqslant m \leqslant M$, on a

$$(\forall (i,j) \in I^{[2]}), \quad \eta_{ij} = \sum \{ \varphi_{D(m)}^{(i)} \varphi_{D(m')}^{(j)} / 1 \leqslant m < m' \leqslant M \}. \quad (5')$$

On commence, comme dans le cas I.1, par introduire l'indice "brut" de proximité

$$s(\omega, \varpi) = \operatorname{card}(R(\omega) \cap R(\varpi)) = \sum \{ \xi_{ij} \eta_{ij} / (i,j) \in I^{[2]} \}$$

$$= \sum \Big\{ \sum \{ \varphi_{C(\ell)}^{(i)} \varphi_{D(m)}^{(i)} \varphi_{C(\ell')}^{(j)} \varphi_{D(m')}^{(j)} / (i,j) \in I^{[2]} \} \Big/ 1 \leqslant \ell < \ell' \leqslant L, \ 1 \leqslant m < m' \leqslant M \Big\}$$

$$= \sum \{ n(\ell \wedge m)\, n(\ell' \wedge m') / 1 \leqslant \ell < \ell' \leqslant L, \ 1 \leqslant m < m' \leqslant M \}. \quad (6)$$

L'h.a.l. de même nature que $N_1$ associe au préordre total $\omega$ (resp. $\varpi$), un préordre total aléatoire $\omega'$ (resp. $\varpi'$) dans l'ensemble, muni d'une probabilité uniforme, $\Omega(n;u)$ (resp. $\Omega(n;v)$) de tous les préordres totaux sur $I$ de même composition (i.e. suite des cardinaux des classes) $u$ (resp. $v$).

A $s(\omega, \varpi)$, nous associons les deux v.a. $s(\omega, \varpi')$ et $s(\omega', \varpi)$ qu'on démontre avoir la même loi [CLERMAN (1973)]. Ainsi, l'expression de $s(\omega, \varpi')$ est

$$s(\omega,\varpi') = \sum \Big\{ \sum \{ \varphi_{C(\ell)}{}^{(i)} \varphi_{C(\ell')}{}^{(j)} \varphi_{D'(m)}{}^{(i)} \varphi_{D'(m')}{}^{(j)} / (i,j) \in I^{[2]} \} / 1 \leqslant \ell < \ell' \leqslant L,\ 1 \leqslant m < m' \leqslant M \Big\}, \quad (7)$$

où $\{ D'(m) / 1 \leqslant m \leqslant M \}$ est la suite des classes du préordre aléatoire $\varpi'$.

Le calcul de la moyenne et de la variance de $s(\omega,\varpi')$ ou de $s(\omega',\varpi)$ (cf. références citées ci-dessus) conduit aux expressions suivantes :

$$\mathcal{E}\big(s(\omega,\varpi')\big) = \lambda\mu$$

$$\mathrm{Var.}\big(s(\omega,\varpi')\big) = \lambda\mu + \rho_{cc}\sigma_{cc} + \rho_{ff}\sigma_{ff} + 2\rho_{cf}\sigma_{cf} + (\theta\varsigma - \lambda^2\mu^2). \quad (9)$$

Les expressions de $\mu$, $\sigma_{cc}$, $\sigma_{ff}$, $\sigma_{cf}$ et $\varsigma$ sont respectivement de même forme que celles de $\lambda$, $\rho_{cc}$, $\rho_{ff}$, $\rho_{cf}$ et $\theta$ ; si les premières sont relatives à la composition $\{ n(\ell) / 1 \leqslant \ell \leqslant L \}$ du préordre $\omega$, les secondes sont relatives à la composition $\{ n(m) / 1 \leqslant m \leqslant M \}$. Plus précisément

$$\lambda = \frac{1}{\sqrt{n(n-1)}} \sum \{ n(\ell)\, n(\ell') / 1 \leqslant \ell < \ell' \leqslant L \}$$

$$\rho_{cc} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{ n(\ell)\, n[c(\ell)]\, (n[c(\ell)] - 1) / 2 \leqslant \ell \leqslant L \}$$

$$\rho_{ff} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{ n(\ell)\, n[f(\ell)]\, (n[f(\ell)] - 1) / 1 \leqslant \ell \leqslant (L-1) \}$$

$$\rho_{cf} = \frac{1}{\sqrt{n(n-1)(n-2)}} \sum \{ n(\ell)\, n[c(\ell)]\, n[f(\ell)] / 2 \leqslant \ell \leqslant (L-1) \}$$

$$\theta = \frac{1}{\sqrt{n(n-1)(n-2)(n-3)}} \sum \Big\{ n(h)\, n(h') \Big[ \sum \{ n(\ell)\, n(\ell') / 1 \leqslant \ell < \ell' \leqslant L \} + n(h) + n(h') - 2n + 1 \Big] \Big/ 1 \leqslant h < h' \leqslant L \Big\} \quad (10)$$

où on note

$n[c(\ell)] = \sum \{ n(\ell') / \ell' < \ell \}$ et $n[f(\ell)] = \sum \{ n(\ell') / \ell' > \ell \}$

L'indice d'association qui généraliserait le coefficient de K. Pearson pour la situation considérée ici, s'écrit

$$\{ s(\omega, \varpi) - \mathcal{E}[s(\omega, \varpi')] \} / \sqrt{\text{var.}[s(\omega, \varpi')]} \quad . \quad (11)$$

La méthode de classification hiérarchique basée sur la vraisemblance des liens nécessite la référence à une échelle $[0, 1]$ de probabilité pour la comparaison deux à deux de l'ensemble des éléments à organiser en classes et sous-classes de proximité. Dans le cas où l'organisation porte sur un ensemble d'échelles définies par des variables qualitatives ordinales, le passage de (11) à un indice de probabilité se justifie par la tendance asymptotiquement normale de la v.a. $s(\omega, \varpi')$ [LERMAN (1976)].

## II - LIAISONS ENTRE LIGNES OU COLONNES D'UNE JUXTAPOSITION DE TABLES DE CONTINGENCE.

### II.1. Structure d'une telle juxtaposition.

La structure la plus simple est celle d'une seule table de contingence

$$\{ n(i \wedge j) / (i, j) \in I \times J \} \quad , \quad (1)$$

où $I$ (resp. $J$) est l'ensemble des codes des modalités d'une variable qualitative nominale de partitionnement. $n(i \wedge j)$ est le nombre d'individus de l'échantillon étudié qui possèdent la $i$-ème modalité du premier caractère et la $j$-ème du second.

Dans le cas où la partition définie par le premier caractère est discrète ; c'est à dire, où $I$ représente l'ensemble des individus, le tableau (1) se réduit à un tableau d'incidence formé de zéros et de uns dont chaque ligne contient exactement un seul 1. C'est un cas particulier du tableau disjonctif-complet que nous allons bientôt définir.

La donnée à laquelle nous nous intéresserons plus particulièrement ici est définie par une juxtaposition « horizontale » de tables de contingences, indéxée par un ensemble de la forme

$$I \times (J^{(1)} \cup J^{(2)} \cup \cdots \cup J^{(\ell)} \cup \cdots \cup J^{(L)}), \qquad (2)$$

où $I$ (resp. $J^{(\ell)}$, $1 \leqslant \ell \leqslant L$) se trouve défini par l'ensemble des modalités d'une variable-partition ; en d'autres termes, $I$ et chaque $J^{(\ell)}$ est un système exhaustif de modalités exclusives.

Nous serons davantage concernés ici par la comparaison, relativement à $I$, des éléments ou de sous-ensembles disjoints de l'ensemble suivant $J$ des modalités :

$$J = J^{(1)} \cup J^{(2)} \cup \cdots \cup J^{(\ell)} \cup \cdots \cup J^{(L)} . \qquad (3)$$

Alors que deux modalités $j_\ell$ et $j'_\ell$ d'un même $J^{(\ell)}$ sont exclusives, il n'en est généralement pas de même de deux modalités $j_\ell$ et $j_{\ell'}$ appartenant respectivement à deux ensembles distincts $J^{(\ell)}$ et $J^{(\ell')}$ $(\ell \neq \ell')$. Nous rappellerons ci-dessous la justification que nous venons de fournir dans un dernier article [LERMAN (1982)] de l'usage d'un même coefficient d'association aussi bien pour comparer $j_\ell$ et $j'_\ell$ que pour comparer $j_\ell$ et $j_{\ell'}$, et ce, à travers $I$.

Il est inutile d'insister sur l'importance pratique d'une telle structure (2) des données qu'on rencontre fréquemment, notamment en Géographie Sociale, Économie, Linguistique, ...

Lorsque la partition indéxant les lignes se réduit à la partition discrète, $I$ représente l'ensemble des individus ou objets et le tableau des données est dans ce cas communément appelé "tableau disjonctif complet". Il s'agit dans ce cas d'un tableau d'incidence où chaque $I \times J^{(\ell)}$ $(1 \leqslant \ell \leqslant L)$ est d'une forme précisée ci-dessus. Un tel tableau se rencontre comme résultat d'un questionnaire où $J^{(\ell)}$ $(1 \leqslant \ell \leqslant L)$ représente l'ensemble des codes des modalités de la $\ell$-ème question.

La structure qui nous paraît la plus générale est celle croisant deux ensembles disjoints de variables qualitatives. L'ensemble d'indexation prend la forme suivante

$$[I^{(1)} \cup I^{(2)} \cup \ldots \cup I^{(k)} \cup \ldots \cup I^{(K)}] \times [J^{(1)} \cup J^{(2)} \cup \ldots \cup J^{(l)} \cup \ldots \cup J^{(L)}]$$

où la restriction du tableau à $I^{(k)} \times J^{(l)}$ est une vraie table de contingence.(4)

Dans la réalité des données, un tel tableau peut se présenter relativement à un questionnaire dont l'ensemble des questions se trouve scindé en deux parties disjointes se référant respectivement à deux aspects de l'étude. Par exemple, dans les enquêtes psychosociologiques, le premier aspect peut être relatif à l'identification sociologique du sujet et le second, à son comportement psychologique.

On avance souvent une structure du tableau des données appelée "tableau de Burt" qui résulte du croisement par lui-même de l'ensemble des modalités des différentes variables qualitatives indexant les colonnes d'un tableau disjonctif complet ; lequel représente la structure essentielle d'origine. Quant au tableau de "Burt", dont la structure propre est intéressante, son rôle dans les différentes méthodes d'analyse des données, est celui d'un intermédiaire utile et pertinent de calcul.

Pour terminer, nous allons considérer la structure d'un tableau définissant une correspondance "floue". Relativement à une classe d'attributs $B$ définissant un profil d'attitude et formée d'attributs orientés (i.e. si $a \leftarrow B \Leftrightarrow \bar{a} \notin B$), nous avions défini dans [LERMAN (1979), (1981a)] le "degré d'appartenance" d'un individu $x$ au type défini par $B$ au moyen de la proportion d'attributs de $B$ possédés par $x$ :

$$\varphi_B(x) = \frac{1}{card(B)} \sum \{\varphi_b(x) / b \leftarrow B\} \quad ,(5)$$

où $\varphi_b(x) = 1$ (resp. 0) selon que l'attribut $b$ est présent (resp. absent) chez $x$.

Soit à présent $\{A_j / 1 \leqslant j \leqslant h\}$ une partition d'un ensemble A d'attributs orientés, obtenue au moyen d'un programme de classification et définissant une typologie de l'ensemble E des individus ou objets. Soit d'autre part c une variable qualitative nominale dont l'ensemble des modalités est noté $\{c_i / 1 \leqslant i \leqslant k\}$, définissant une partition nette sur l'ensemble des individus ou objets. Nous associons dans ces conditions un tableau $k \times h$ définissant une correspondance "floue" et croisant les modalités $c_i$, $1 \leqslant i \leqslant k$, de la variable c avec les classes $A_j$, $1 \leqslant j \leqslant h$, de la typologie. Pour ce tableau qu'on notera

$$\{ \nu(i,j) / 1 \leqslant i \leqslant k, 1 \leqslant j \leqslant h \}, \quad (6)$$

$\nu_{ij}$ qui se trouve défini comme suit

$$\nu_{ij} = \sum \{ \varphi_{c_i}(x) \varphi_{A_j}(x) / x \in E \} \quad (7)$$

mesure l'importance numérique de la classe "floue" résultant de l'intersection de la classe "nette" définie par la modalité $c_i$ et de la classe "floue" définie par l'ensemble $A_j$ des attributs.

$\{A_i / 1 \leqslant i \leqslant k\}$ et $\{B_j / 1 \leqslant j \leqslant h\}$ définissant respectivement deux partitions de deux ensembles disjoints A et B formés d'attributs orientés, la correspondance la plus générale [LERMAN (1979), (1981a)] est celle croisant les deux typologies définies par les deux précédentes partitions. Le contenu de la case $(i,j)$ du tableau $k \times h$ de la correspondance se trouve défini par

$$\nu_{ij} = \sum \{ \varphi_{A_i}(x) \varphi_{B_j}(x) / x \in E \}. \quad (8)$$

qui mesure l'importance numérique de la classe "floue" résultant de l'intersection des deux classes "floues" respectivement définies par les deux ensembles $A_i$ et $B_j$ d'attributs.

## II.2 - Indice d'association entre colonnes ou lignes d'une juxtaposition de tables de contingences.

### II.2.1 - Cas d'un seul tableau de contingence.

#### a) Représentation géométrique.

Reprenons le tableau (1) de contingence ci-dessus :

$$\{ n_{ij} \,/\, (i,j) \in I \times J \} \qquad (9)$$

où $n_{ij}$ désigne le nombre d'individus possédant les modalités $i$ de $I$ et $j$ de $J$. On rappelle les notations :

$$\left.\begin{aligned} n(i) = \sum \{ n_{ij} \,/\, j \in J \},\ n(j) = \sum \{ n_{ij} \,/\, i \in I \} \\ n = \sum \{ n(i) \,/\, i \in I \} = \sum \{ n(j) \,/\, j \in J \} \end{aligned}\right\} \qquad (10)$$

Le principe de la comparaison de deux modalités $j$ et $j'$ de $J$, repose sur la représentation euclidienne de $I$ à travers $J$ [LERMAN, TALLUR (1980)]. Cette représentation est fournie dans le cadre de l'analyse des correspondances où on associe à chaque $i$ de $I$ le point de $\mathbb{R}^{|J|}$ dont la suite des coordonnées est $\{ n_{ij}/n(i) \,/\, j \in J \}$; il s'agit du « profil » de $i$ à travers $J$.

Dès lors, chaque $j$ de $J$ se trouve assimilé à une variable quantitative dont la valeur sur le $i$-ème point est égale à $f_j^i = n_{ij}/n(i)$, ce dernier étant affecté du poids $p(i) = n(i)/n$, $i \in I$. Dans ces conditions, le coefficient de corrélation entre $j$ et $j'$ de $J$, se met sous la forme :

$$\rho(j,j') = \frac{\sum \{ p(i) [f_j^i - p(j)][f_{j'}^i - p(j')] \,/\, i \in I \}}{\sqrt{\sum_i p(i) [f_j^i - p(j)]^2 \sum_i p(i) [f_{j'}^i - p(j')]^2}} \qquad (11)$$

Pour obtenir l'indice d'association entre deux éléments $i$ et $i'$ de $I$, on transposera la situation en considérant la représentation euclidienne de $J$ à travers $I$.

b) Représentation ensembliste.

Désignons par $\{E_i / i \in I\}$ (resp. $\{F_j / j \in J\}$) la partition sur l'ensemble $E$ des objets, définie par la variable qualitative nominale dont les modalités indexent les lignes (resp. colonnes) du tableau de contingence.

Si $j$ et $h$ sont les deux modalités exclusives de $J$ à comparer relativement à $I$, la situation peut être, par rapport à un même $i$, schématisée comme suit :

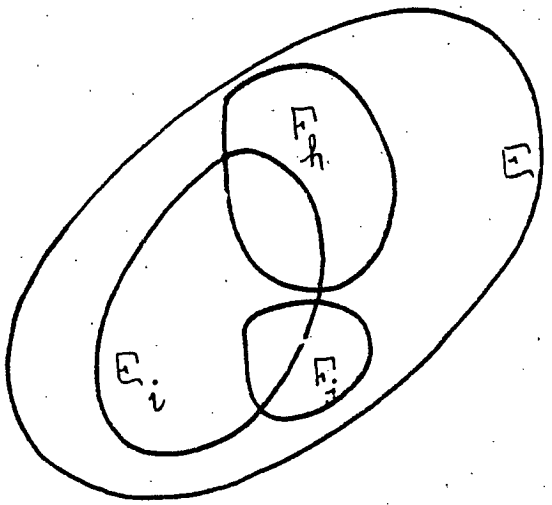



Fig. 2

Nous allons donner un bref aperçu sur l'élaboration d'un indice conforme au schéma de construction rappelé au paragraphe I ci-dessus et dont l'analyse complète se trouve dans [LERMAN (1982)].

b.1. Indice brut d'association.

Relativement à la classe $E_i$, le lien brut entre les deux parties $F_j$ et $F_h$ sera mesuré par

$$\varepsilon(j,h/i) = \frac{n(i\wedge j)\, n(i\wedge h)}{n(i)} = n(i)\, \frac{n(i\wedge j)}{n(i)} \times \frac{n(i\wedge h)}{n(i)} \quad . \quad (12)$$

Dans ces conditions, le lien brut entre $F_j$ et $F_h$, relativement à la partition $\{E_i / i \in I\}$, se trouve défini par

$$\sum \left\{ \frac{n(i \wedge j)\, n(i \wedge h)}{n(i)} \Big/ i \in I \right\} = \sum \left\{ n(i) \frac{n(i \wedge j)}{n(i)} \times \frac{n(i \wedge h)}{n(i)} \Big/ i \in I \right\},$$

avec des notations que l'on comprend. (13)

b.2. L'hypothèse d'absence de lien.

Une forme de l'h.a.l. de même nature que $N_1$ (cf. § I) consiste à fixer $F_j$ et $F_h$ et à associer à la partition $\{E_i / 1 \leqslant i \leqslant |I|\}$, une partition aléatoire $\{X_i / 1 \leqslant i \leqslant |I|\}$ dans l'ensemble $\mathcal{P}(n;t)$, muni d'une probabilité uniforme, des partitions en classes étiquetées de même type $t = [n(1), n(2), \ldots, n(i), \ldots, n(|I|)]$.

La forme duale de l'h.a.l. consiste à fixer la partition $\{E_i / 1 \leqslant i \leqslant |I|\}$ et à associer à la partition $\{F_j / 1 \leqslant j \leqslant |J|\}$, une partition aléatoire $\{Y_j / 1 \leqslant j \leqslant |J|\}$, en classes étiquetées et de même type $s = [n(1), n(2), \ldots, n(j), \ldots, n(|J|)]$, dans l'ensemble $\mathcal{P}(n;s)$, muni d'une probabilité uniforme, des partitions de $E$, en classes étiquetées de type $s$.

A l'indice brut $s(j,h/P)$ défini par la formule (13) ci-dessus, on associe les deux v.a. duales

$$S(j,h/P) = \sum \left\{ \frac{\operatorname{card}(X_i \cap F_j)\, \operatorname{card}(X_i \cap F_h)}{n(i)} \Big/ 1 \leqslant i \leqslant |I| \right\} \qquad (14)$$

et

$$T(j,h/P) = \sum \left\{ \frac{\operatorname{card}(E_i \cap Y_j)\, \operatorname{card}(E_i \cap Y_h)}{n(i)} \Big/ 1 \leqslant i \leqslant |I| \right\} \qquad (15)$$

Nous démontrons [LERMAN (1982)] que ces deux v.a. duales ont la même répartition. Nous allons donner les expressions de la moyenne et de la variance de la distribution commune dont les calculs sont détaillées dans la référence citée.

b.3- Moyenne et Variance de $S(j,h/P)$ (resp. $T(j,h/P)$).

La moyenne et la variance sont respectivement égales à

$$\frac{(n-|I|)}{(n-1)} \times \frac{n(j)\,n(h)}{n} \qquad (16)$$

et

$$\sum_{1\leqslant i\leqslant |I|} \frac{[n(i)-1]}{n(i)} \times \frac{n(j)\,n(h)}{n(n-1)} \left\{ 1 + \frac{[n(i)-2][n(h)-1]}{(n-2)} + \frac{[n(i)-2][n(j)-1]}{(n-2)} \right.$$

$$\left. + \frac{[n(i)-2][n(j)-1]}{(n-2)} + \frac{[n(i)-2][n(i)-3][n(j)-1][n(h)-1]}{(n-2)(n-3)} \right\}$$

$$+ 2 \sum_{1\leqslant i<i'\leqslant |I|} \frac{[n(i)-1][n(i')-1]\,n(j)[n(j)-1]\,n(h)[n(h)-1]}{n(n-1)(n-2)(n-3)}$$

$$- \left[\frac{n-|I|}{n-1}\right]^2 \times \left[\frac{n(j)n(h)}{n}\right]^2 . \qquad (17)$$

Il ne faut pas s'étonner de constater que la moyenne (16) est nulle si la partition $P = \{E_i / i \in I\}$ est discrète ; en effet, dans ce cas, l'indice brut qui se réduit à $n(j \wedge h)$ est lui même nul puisque $j$ et $h$ sont deux modalités exclusives. Généralement, pour une table de contingence courante, $|I|$ est «petit» devant $n$, de sorte que l'expression (16) peut être approchée par $n(j)n(h)/n$. En admettant cette approximation, on peut se rendre compte que le numérateur de l'expression (11) ci-dessus est, au facteur $1/n$ près, l'indice centré

$$\sum_{1\leqslant i\leqslant |I|} \frac{n(i\wedge j)n(i\wedge h)}{n(i)} - \frac{n(j)n(h)}{n} \qquad (18)$$

Toutefois, le carré du dénominateur de l'indice de corrélation (11) a une expression essentiellement différente de la variance (17).

L'indice que nous obtenons en centrant et en réduisant l'indice brut $S(j,h/P)$ est donc de nature nouvelle par rapport à l'indice

de corrélation (11).

II.2.2- Cas d'une juxtaposition «horizontale» de tables de contingences.

Nous avons ci-dessus précisé la structure d'un tel tableau de données dont l'ensemble d'indexation est défini par l'expression (2) du paragraphe II.1:

$$I \times ( J^{(1)} \cup J^{(2)} \cup \ldots \cup J^{(l)} \cup \ldots \cup J^{(L)} ) \quad . (19)$$

Comme nous l'avons déjà signalé, nous nous intéresserons surtout à la définition d'un indice d'association entre éléments de l'ensemble J des modalités :

$$J = J^{(1)} \cup J^{(2)} \cup \ldots \cup J^{(l)} \cup \ldots \cup J^{(L)} \qquad (20)$$

Dans [LERMAN, TALLUR (1980)], nous indiquons comment nous utilisons l'indice de corrélation (11) ci-dessus, que les modalités j et j' à comparer se réfèrent à un même $J^{(l)}$ ou non. Dans le premier cas l'indice est conçu au niveau de la seule table de contingence $I \times J^{(l)}$ et dans le second, où $(j, j') \in J^{(l)} \times J^{(l')}$ avec $l \neq l'$, l'indice est calculé au niveau de la juxtaposition d'exactement deux tables de contingence, indexée par $I \times (J^{(l)} \cup J^{(l')})$. Alors que dans la méthode de l'Analyse des Correspondances Multiple, on doit assimiler la totalité de la juxtaposition à une seule table de contingence [LEBART & al. (1977)].

Relativement au point de vue de la représentation ensembliste des variables, nous avons considérée dans [LERMAN (1982)] une h.a.l. plus libre que celle considérée ci-dessus (§ b.2) et qui intègre aussi bien le cas où deux modalités j et h sont exclusives que celui où elles ne le sont pas.

a) Indice brut et h.a.l.

La situation peut ici être schématisée par la figure suivante où, à la différence de la figure 1 précédente, les parties $F_j$ et $F_h$ peuvent avoir une intersection non vide. On adopte la même

$$\text{var}[U(j,h/P)] = e(j,h)\Big\{\sum_{1\le i\le |I|}\Big\{\frac{[n(i)-1]^2[n(j)-1][n(h)-1]}{n(n-1)^2} + \frac{[n(i)-1][n(j)-1]}{n(n-1)} + \frac{[n(i)-1][n(h)-1]}{n(n-1)} + \frac{1}{n}\Big\} + 2\sum_{1\le i<i'\le |I|}\frac{n(i)\,n(i')[n(j)-1][n(h)-1]}{[n(n-1)]^2} - e(j,h)\Big\}, \quad (23)$$

où $e(j,h)$ désigne le paramètre $n(j)n(h)/n$. On a

$$\text{var}[U(j,h/P)] \simeq e(j,h)\{p(j)+p(h)+(|I|/n)\}, \quad (24)$$

où $p(j)$ (resp. $p(h)$) désigne la proportion $n(j)/n$ (resp. $n(h)/n$).

Comme nous l'avons déjà constaté, l'indice centré est exactement, au coefficient $n$ près, le numérateur de l'indice de corrélation $\rho(j,h)$ défini par la formule (11) ci-dessus.

II.2.2.1 - Cas où I définit une partition discrète.

Il s'agit du cas d'un tableau "disjonctif complet" mentionné au paragraphe II.1 ci-dessus. On peut se rendre compte que dans ce cas, l'indice centré réduit

$$\frac{s(j,h/P) - \mathcal{E}[U(j,h/P)]}{\sqrt{\text{var}[U(j,h/P)]}} \quad (25)$$

est exactement, au coefficient $\sqrt{n-1}$ près, l'indice d'association de K. Pearson entre les deux attributs-modalités $j$ et $h$.

Il est important de noter que pour cette situation, l'indice de corrélation $\rho(j,h)$ ((11) ci-dessus) se réduit également à celui de K. Pearson.

Résumons nous au moyen de l'énoncé suivant :

Théorème. Dans le cadre de l'h.a.l. ci-dessus définie, l'indice $s(j,h/P)$ centré (numérateur de (25)) est, au coefficient $n$ près, égal au numérateur de l'indice de corrélation $\rho(j,h)$. Le dénominateur de l'indice centré réduit (25) reste de forme essentiellement différente du dénominateur de $\rho(j,h)$. Toutefois, les deux indices ont la même forme, celle du coefficient de K. Pearson, dans le cas d'un tableau disjonctif complet où la partition $\{E(i)/i \in I\}$ est discrète.

Indépendamment de ce résultat qui nous conforte, nous avons l'habitude d'utiliser comme indice d'association entre attributs-modalités, le même indice que celui entre attributs quelconques. L'analyse expérimentale telle que celle de [LERMAN (1981a), Chap. 11], nous a conduit à préférer la forme $N_3$ de l'h.a.l. (cf. § I.1) pour laquelle l'indice d'association entre les modalités $j$ et $h$ prend la forme

$$Q(j,h) = \frac{\sqrt{n}\,[\,p(j\wedge h) - p(j)\,p(h)\,]}{\sqrt{p(j)\,p(h)}}, \qquad (26)$$

où $p(j\wedge h)$, $p(j)$ et $p(h)$ représentent les proportions $n(j\wedge h)/n$, $n(j)/n$ et $n(h)/n$.

Désignons par $j_\ell$ une modalité courante de l'ensemble $J^{(\ell)}$ des modalités exclusives et exhaustives de la $\ell$-ème variable qualitative.

Dans une classification automatique par proximité de l'ensemble $J$ ($J = J^{(1)} \cup \dots \cup J^{(\ell)} \cup \dots \cup J^{(L)}$) des modalités réunies, on peut voir apparaître dans une même classe plus d'une seule modalité d'un même $J^{(\ell)}$. Il faut savoir que cette association entre modalités exclusives se fait nécessairement par enchaînements successifs à travers une suite de modalités dont deux consécutives sont non exclusives et proches.

En effet, pour $\ell \neq \ell'$, $j_\ell$, $h_\ell \in J^{(\ell)}$ et $j'_{\ell'} \in J^{(\ell')}$, il

suffit que la densité $p(j_\ell \wedge j'_{\ell'}) / p(j_\ell)\, p(j'_{\ell'})$ soit supérieure (strictement) à $\left(1 - \sqrt{p(h_\ell)/p(j'_{\ell'})}\right)$ pour avoir

$$\varphi(j_\ell, h_\ell) < \varphi(j_\ell, j'_{\ell'})$$

On a même de plus, pour tous $j_\ell$ et $\ell'$ $(\ell' \neq \ell)$ fixés,

$$p(j_\ell \wedge j'_{\ell'}) / p(j_\ell)\, p(j'_{\ell'}) \gg 1$$

pour au moins un $j'_{\ell'}$ de $J^{(\ell')}$. On montrera en effet, que dans le cas contraire, c'est impossible.

### II.2.3- Cas d'une correspondance "floue"

La structure d'une telle correspondance a été définie à la fin du paragraphe II.1 auquel on se reportera pour les notations. L'indice d'association entre éléments de J (resp. de I) peut correspondre à l'un des deux indices (11) ou (25) qui ne diffèrent d'ailleurs qu'au niveau du dénominateur.

Dans chacun des cas, on remplacera les entiers $n(i \wedge j)$ par les nombres rationnels positifs $\nu(i,j)$ (cf. selon le cas formule (7) ou (8) § II.1).

Signalons qu'on a obtenu [J.L. BUARD (1980)] d'excellents résultats concrets pour ce type de données. Il s'agissait, dans le cadre du centre hospitalier régional de Rennes, de typer la demande d'actes par nature faites par les unités fonctionnelles (ou unités de soins) dans les différents laboratoires (Biochimie, Bactériologie, etc ...) et puis, surtout, de déterminer une typologie de l'ensemble des unités fonctionnelles, relativement aux différentes classes d'actes.

Dans cette étude, l'unité statistique est la "classe" qui

prévoit sous la forme d'un questionnaire (avec des cases à cocher), la suite des actes de laboratoire à effectuer. Il faut savoir que pour la seule année de novembre 77 à octobre 78 et la seule rubrique de "Biochimie" - qui seule a été traitée pendant la période mentionnée - il y a environ 135000 liasses.

J.L. Buard a directement extrait le tableau carré symétrique des indices bruts de proximité entre actes demandés, lesquels étant assimilés à des attributs de description (cf. § I.1). Après référence à une échelle de probabilité, conformément à l'h.a.l. $N_3$, pour la comparaison deux à deux, l'application de l'Algorithme de la Vraisemblance du Lien - dont nous rappellerons ci-dessous le principe - n'a nullement été sensible, à l'inverse des méthodes de classification classiques, à l'hétérogénéité de la fréquence de demande des différents actes.

On détecte une partition $\{B_j / 1 \leq j \leq |J|\}$ de l'ensemble des attributs (ici des "actes") à un niveau significatif de l'arbre des classification [LERMAN (1981a), Chap. 4 et 5, BUARD (1980)]. Le croisement de l'ensemble des unités de soin que nous représentons par I, avec les classes $B_j$ $(1 \leq j \leq |J|)$ d'attributs, conduit à une correspondance floue où $\nu_{ij}$ $(1 \leq i \leq |I|, 1 \leq j \leq |J|)$ a la forme (7) du paragraphe I.1.

La figure suivante fournit l'image de l'arbre condensé de classification sur l'ensemble I, obtenu à partir de la description définie par le tableau précédent. L'indice d'association est celui qui correspond à l'indice de corrélation (11). Après référence à une échelle de probabilité à partir de la fonction de répartition de la loi normale [LERMAN (1981), Chap. 2], l'algorithme de classification ascendante hiérarchique utilisé est celui de la vraisemblance du lien (A.V.L.). La réduction de l'arbre se fait aux niveaux où apparaît un "nœud significatif" [LERMAN (1981a), Chap. 4 et 5].

L'interprétation par rapport aux classes $B_j$ $(1 \leq j \leq J)$, d'une partition $\{I_k / 1 \leq k \leq K\}$ prise à un niveau "significatif" de

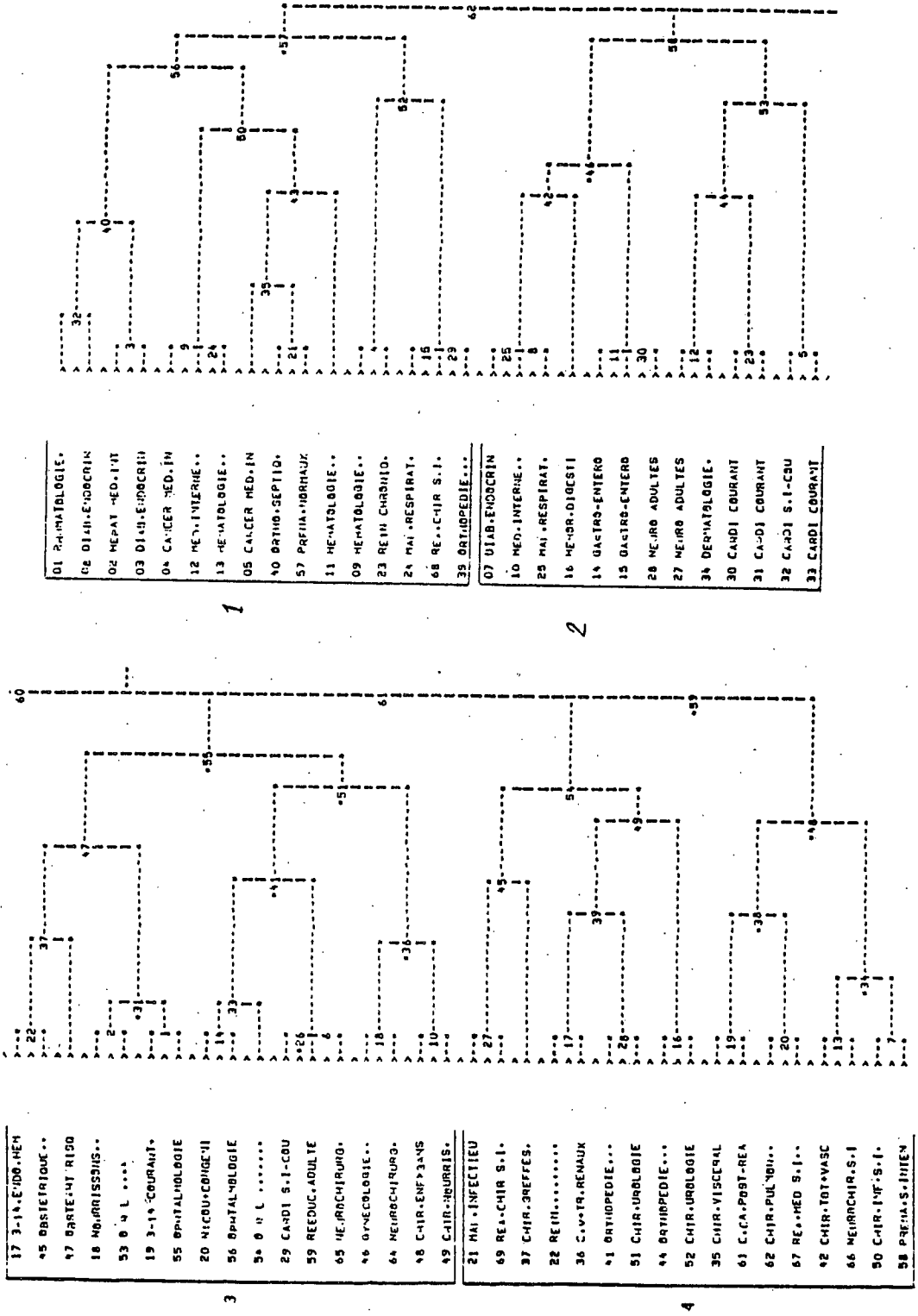

01 RHUMATOLOGIE.
02 DIAB.ENDOCRIN
03 DIAB.ENDOCRIN
04 CANCER MED.IN
05 CANCER MED.IN
07 DIAB.ENDOCRIN
10 MED.INTERNE..
11 HEMATOLOGIE..
12 MED.INTERNE..
13 HEMATOLOGIE..
14 GASTRO-ENTERO
15 GASTRO-ENTERO
16 MEDEC.DIGEST
17 3-14.ENDO.NEM
18 MD.MAISSONS..
19 3-14 COURANT
20 NICOU-CONGEVI
21 MAL.INFECTIEU
22 REIN.....
23 REIN CHRONIO.
24 MAL.RESPIRAT.
26 REEDUC.ADULTE
26 NEURO ADULTES
27 NEURO ADULTES
29 CARDI S.I-COU
30 CARDI COURANT
31 CARDIO COURANT
32 CARDI S.I-COU
33 CARDI COURANT
34 DERMATOLOGIE.
35 CHIR.VISCERAL
36 CHIR.RENAUX
37 CHIR.GREFFES.
38 NEUROCHIRURG.
39 ORTHOPEDIE...
41 ORTHOPEDIE...
44 ORTHOPEDIE...
45 OBSTETRIQUE..
46 GYNECOLOGIE..
48 CHIR.ENFANTS
49 CHIR.NEURIS.
51 CHIR.UROLOGIE
52 CHIR.UROLOGIE
53 0 N L .....
55 RHUMATOLOGIE
56 RHUMATOLOGIE
57 RHUMATOLOGIE.
58 RHUMATOLOGIEN
59 REEDUC.ADULTE
60
61
62 CHIR.PULMON...
63 CHIR.POST-REA
65 NEURO CHIRURG.
66 NEUROCHIR.S.I
67 REA.MED S.I..
68 REA.CHIR S.I.
69 REA.CHIR S.I.
90 CHIR.INF.S.I.
1
2
3
4

Fig. 4.

l'arbre des classifications sur I, se fait en construisant une table de dimension K x J où à l'intersection de la k-ème ligne et de la j-ème colonne, on place la valeur de la contribution orientée à la statistique du $\chi^2$ [LERMAN (1979) repris dans (1981a)] :

$$\frac{\nu(k \wedge j) - [\nu(k)\nu(j)/n]}{\sqrt{\nu(k)\nu(j)/n}}, \quad (27)$$

$1 \leqslant k \leqslant K$, $1 \leqslant j \leqslant J$, dont nous laissons le soin au lecteur de préciser les éléments en se référant à la formule (7) du paragraphe I.1.

Un dernier point pour clore ce paragraphe concernera, dans le cas d'une juxtaposition de tables de contingences, la définition d'un indice d'association entre deux éléments $i$ et $i'$ de I.

Considérons pour fixer les idées, le cas d'une juxtaposition « horizontale » de tables de contingence indexée par $I \times (J^{(1)} \cup J^{(2)} \cup \cdots \cup J^{(\ell)} \cup \cdots \cup J^{(L)})$. Deux points de vue peuvent prévaloir pour la comparaison de $i$ et $i'$.

Pour le premier, on commence par réduire chaque table $I \times J^{(\ell)}$, $1 \leqslant \ell \leqslant L$, à une distribution de fréquences $\{f_{i j_\ell} / (i, j_\ell) \in I \times J^{(\ell)}\}$ de somme 1, puis, en divisant par L chaque $f_{i j_\ell}$ pour tout $\ell$, on se ramène à une distribution sur $I \times J$ de nombres positifs de somme 1. C'est l'attitude adoptée dans l'analyse factorielle des correspondances multiples [LEBART & al. (1977)].

Pour le deuxième point de vue, on calcule la contribution de chaque $J^{(\ell)}$, $1 \leqslant \ell \leqslant L$, sur la base de la table $I \times J^{(\ell)}$, à l'association entre $i$ et $i'$, puis on prend une moyenne (équipondérée ou non) des différentes contributions.

L'un ou l'autre des indices développés aux paragraphes II.2.1 ou II.2.2. peut être adapté pour l'un ou l'autre des deux

points de vue. Toutefois, l'indice de corrélation semble davantage adapté au premier point de vue et ceux résultant d'une représentation ensembliste, au second point de vue.

## III- FORMATION ASCENDANTE HIERARCHIQUE DE L'ARBRE DES CLASSIFICATIONS.

### III.1- Algorithme de la Vraisemblance du Lien (A.V.L.).

L'indice d'association, que nous allons rapidement rappeler, entre parties disjointes de l'ensemble à classifier est utilisé dans notre méthode pour organiser en classes et sous-classes de proximité aussi bien l'ensemble des lignes que celui des colonnes d'un tableau des données, quelle que soit sa nature mathématique.

Désignons par B l'ensemble à classifier. On suppose qu'on dispose de la table des proximités entre éléments de B :

$$\{ P(c,d) / \{c,d\} \in P_2(B) \} \qquad (1)$$

où $P_2(B)$ est l'ensemble des paires ou parties à deux éléments de B et où $P(c,d)$ se réfère à une échelle de probabilité définie par la vraisemblance du lien, pour tout $\{c,d\}$ de $P_2(B)$. Si par exemple, B se trouve formé d'attributs descriptifs, $P(c,d)$ peut être défini par $Pr\{ card(X \cap Y) < card[I(c) \cap I(d)] / N_i \}$ où $(X,Y)$ est le couple de parties aléatoires indépendantes, associé dans l'h.a.l. $N_i$, au couple $(I(c), I(d))$ (cf. § I.1).

Plus particulièrement, pour le cas qui nous concerne de la classification d'un ensemble J de modalités, ayant abouti à l'un des indices ci-dessus présentés (cf. § II.2.1 et II.2.2), que nous noterons

$$\{ Q(j,h) / \{j,h\} \in P_2(J) \} \qquad (2)$$

nous commençons par centrer et réduire globalement - sur $P_2(J)$ - cette variable "proximité", pour aboutir à la table

$$\{ Q'(j,h) \,/\, \{j,h\} \in P_2(J) \} \qquad (3)$$

avec

$$\left(\forall \{j,h\} \in P_2(J)\right),\ Q'(j,h) = [Q(j,h) - \text{moy.}(Q)] \,/\, \sqrt{\text{var.}(Q)}\ , \qquad (4)$$

où moy.(Q) et var.(Q) sont respectivement la moyenne et la variance de la distribution (2).

Non sans l'avoir dans une certaine mesure méthodologiquement justifié, on aboutit à la table suivante de même structure que celle (1),

$$\{ P(j,h) \,/\, \{j,h\} \in P_2(J) \}\ , \qquad (5)$$

en posant $P(j,h) = \Phi(Q'(j,h))$, pour tout $\{j,h\} \in P_2(J)$, où $\Phi$ désigne la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$.

Revenons au cas général et désignons par C et D deux parties disjointes de B. Le point de départ de l'élaboration de l'indice de comparaison entre C et D est fourni, à partir de considérations topologiques par la formule

$$p(C,D) = \max \{ P(c,d) \,/\, (c,d) \in C \times D \}\ . \qquad (6)$$

L'indice final que nous retenons résulte de la distribution de la v.a. $p(C',D')$ associée dans l'h.a.l. à (6) où C' (resp. D') est une classe aléatoire associée à C (resp. D). Cet indice auquel on aboutit prend la forme très simple suivante :

$$P(C,D) = [p(C,D)]^{\ell \times m}\ , \qquad (7)$$

où $\ell = \text{card}(C)$ et $m = \text{card}(D)$.

Considérons l'arbre hiérarchique « détaillé » des classifications, obtenu pas à pas, où à chaque pas on réunit la paire de classes (resp. les paires de classes s'il y en a plus

d'une) qui réalise la plus grande valeur de la proximité (7). On peut aisément montrer que cette valeur maximale est fortement décroissante d'un niveau à celui consécutif de l'arbre qui est donc sans inversions.

D'ailleurs, pour le calcul, compte tenu du fait que seule une échelle ordinale suffit pour les comparaisons, on considère $\left(-\text{Log}\{1-\text{Log}[P(C,D)]\}\right)$.

## III.2. Algorithme Basé sur la Corrélation (A.B.C.).

L'étude du paragraphe II nous a permis de nous rendre compte qu'il y a deux approches de la notion de corrélation ; la plus classique se réfère à une représentation géométrique des variables, alors que celle, que nous avons particulièrement développée, se réfère à une représentation ensembliste des variables.

Quelle que soit la conception retenue de l'indice, ce dernier peut être utilisé pour former la totalité de l'arbre binaire des classifications sur l'ensemble J des modalités selon l'algorithme classique de classification ascendante hiérarchique. En effet, la fusion des deux modalités j et h les plus voisines, conduit à la création d'une nouvelle modalité $j \vee h$ ($j$ ou $h$) et on se retrouve à chaque étape à rechercher dans l'ensemble des nouvelles modalités, les deux les plus proches. Certes, il peut se faire qu'à un même niveau de la construction, on trouve plusieurs paires de modalités également les plus proches ; dans ce cas, on procède - dans un ordre quelconque - à la suite des agrégations des paires les plus voisines et les classes résultantes sont placées à un même niveau de l'arbre.

C'est cette idée qui a conduit B. Tallur à construire un algorithme de classification ascendante hiérarchique qui utilise uniquement l'indice de corrélation $P(j,h)$ (cf. formule (11) § II) et qu'il intitule du titre de ce paragraphe [TALLUR (1982)].

Pour que d'un niveau au suivant l'arbre soit sans inversion, B. Tallur montre qu'il importe d'apporter un coefficient correctif et d'adopter comme indice d'association entre la nouvelle agrégation $j \vee h$ et $k$ $(k \neq j, k \neq h)$, $\frac{1}{\sqrt{2}} p(j \vee h, k)$, où $k$ est une classe de modalités existant au dernier niveau formé de l'arbre.

Un tel cœfficient $(1/\sqrt{2})$ qu'on peut chercher à justifier de façon plus précise, se comprend très intuitivement si on veut préserver le caractère entier des unités statistiques initiales que sont les modalités de $J$. De toute façon, de la sorte, l'algorithme donne d'excellents résultats.

L'indice $Q(j;h)$ auquel nous sommes parvenus au paragraphe II.2.2 et qui ne diffère de $P(j,h)$ qu'au niveau du dénominateur, a pour expression explicite

$$Q(j;h) = \frac{\sum_{1 \leq i \leq n} \frac{n(i \wedge j)\, n(i \wedge h)}{n(i)} - \frac{n(j)\, n(h)}{n}}{\sqrt{n\, p(j)\, p(h) \left[ p(j) + p(h) + \iota \right]}}, \qquad (8)$$

où $\iota = |I|/n$.

Nous démontrons dans [LERMAN (1982)] que c'est le même cœfficient correctif $1/\sqrt{2}$ qui est nécessaire pour s'assurer le passage sans inversion d'un niveau de l'arbre à celui, consécutif. De façon beaucoup plus générale nous y démontrons le résultat suivant

Théorème. L'indice d'association entre deux classes disjointes $G$ et $H$ de modalités de $J$ ($\ell = \mathrm{card}(G)$, $m = \mathrm{card}(H)$) qui généralise l'indice $Q(j,h)$ ci-dessus, pour l'obtention par l'algorithme de classification hiérarchique ascendante d'un arbre binaire sans inversions est défini par

$$Q_c(G,H) = \left(\frac{1}{\sqrt{2}}\right)^{(\ell + m - 1)} Q(G,H), \qquad (9)$$

où $Q(G,H)$ est l'indice (8) appliqué à deux colonnes sommes ;

la première (resp. la seconde) résultant de la somme des colonnes des modalités initiales de G (resp. de H) :

$$Q(G,H) = Q\left(\sum_{g\in G}\{n(i\wedge g)/g\in G\}, \sum_{h\in H}\{n(i\wedge h)/h\in H\}\right) \quad (10)$$

III.3. Nœuds significatifs ; condensation de l'arbre.

Revenons au cas général de la classification d'un ensemble B (cf. § III.1). Une étape décisive de notre méthode de classification hiérarchique consiste à condenser l'arbre aux niveaux où se produit un nœud « significatif » détecté à partir du comportement d'une statistique de proximité entre une certaine forme de l'information quant aux ressemblances entre éléments de l'ensemble B et l'association entre deux classes correspondante au nœud.

Le principe d'élaboration de cette statistique de proximité est toujours le même (cf. § I). Mais pour se ramener à la comparaison de deux structures de même type, on ne retient de l'indice de proximité Q sur B que le préordre total associé sur l'ensemble $F = P_2(B)$ des paires d'éléments distincts de B, c'est à dire la "préordonnance" sur B, $\omega(B)$ où

$$(\forall (p,q)\in F\times F),\ p<q \iff Q(p) > Q(q) \quad . \quad (11)$$

En effet, la donnée d'une partition $\pi$, éventuellement produite à un niveau de l'arbre, peut être regardée comme définissant un préordre total sur F à deux classes $R(\pi)$ et $S(\pi)$ où $R(\pi)$ (resp. $S(\pi)$) est l'ensemble des paires réunies (resp. séparées) par la partition $\pi$. $R(\pi) < S(\pi)$ pour l'ordre quotient.

Nous représentons dans $F\times F$, $\omega(B)$ par son graphe :

$$gr(\omega) = \{(p,q)/(p,q)\in F\times F,\ p<q \text{ et non } q<p \text{ pour } \omega\} \quad (12)$$

et la partition $\pi$ par le rectangle $R(\pi)\times S(\pi)$.

L'indice brut entre la préordonnance et la partition sera dans ces conditions

$$s(\omega,\pi) = \text{card}\,[\,gr(\omega) \cap (R(\pi) \times S(\pi))\,] \qquad (13)$$

Le critère a été introduit par J.P. Benzecri [BENZECRI (1965)] sous la forme encore par trop métrique du "nombre d'inégalités entre les distances spécifiées par la partition et compatibles avec l'ordonnance". Nous l'avons repris sous la forme (13) et surtout nous avons étudié sa distribution lorsque $\pi$ est un élément aléatoire dans l'ensemble, muni d'une probabilité uniforme, des partitions de même type (i.e. dont les cardinaux des classes sont fixées [LERMAN (1973), (1981a), (1982a)]. Nous démontrons que cette distribution est asymptotiquement normale et nous caractérisons de façon exacte l'expression formelle de chacun de ses moments.

La statistique "globale" $\mathcal{E}$, obtenue en centrant et en réduisant (13) par rapport à l'h.a.l. estimée, définit la «mesure» d'adéquation globale de la partition. La suite des valeurs observées de $\mathcal{E}$ sur la suite des niveaux de l'arbre des classifications permet une interprétation dynamique de ce dernier et sa condensation aux niveaux où se produit un nœud «significatif». En attachant à chaque niveau i le taux d'accroissement $\theta_i = (\mathcal{E}_i - \mathcal{E}_{i-1})$, un tel nœud apparaît comme un maximum local de la distribution observée de $\theta$ le long de la suite des niveaux de l'arbre. L'examen conjoint de cette dernière distribution et de celle correspondante de $\mathcal{E}$, permet de reconnaître les principaux états d'équilibre dans la synthèse et les principales associations significatives pour atteindre un même état d'équilibre.

## III.4- Contrainte de contiguïté.

Lorsque I est formé d'un ensemble d'unités géographiques (e.g. communes, départements, ...) d'une région donnée, la

problème se pose d'établir une carte de la région qui respecte au mieux les proximités de comportement statistique; mais où chaque zone se trouve nécessairement définie par une partie connexe de I.

A cette fin et sous forme optionnelle, une contrainte de contiguïté a été introduite dans nos programmes, sous la forme d'une table unidimensionnelle attachant à chaque i, le sous ensemble des i' de code inférieur à celui de i et qui «touchent» i [A. PROD'HOMME (1980)]. De la sorte, une association entre deux agrégats ne peut se faire que s'il y a deux éléments, appartenant respectivement aux deux agrégats, qui soient contigüs (i.e. tel que l'un appartienne à la table de contiguïté de l'autre). La carte est alors obtenue, sur la base de la statistique "globale" $C_1$, à un des niveaux les plus significatifs de l'arbre des classifications sur I.

Il peut se faire [B. TALLUR (1978)] que la seule structure de proximité statistique entraîne à elle seule une «bonne part» de la connexité spatiale. Il serait intéressant de pouvoir quantifier ce phénomène pour pouvoir préciser le degré d'entraînement de la connexité spatiale dû aux seules variables de comportement statistique.

Ce même type de programme, où il y a lieu de remplacer la contrainte de contiguïté par une contrainte de discontiguïté, peut de façon intéressante être utilisé pour la classification des lignes (resp. colonnes) d'un tableau d'incidence "creux" (i.e. densité des "1" faible) où on s'interdit l'agrégation de deux classes, s'il n'y a pas deux éléments appartenant respectivement aux deux classes et dont le nombre d'associations positives (i.e. indice brut de proximité) est supérieur à un seuil donné [J.L. BUARD (1980)].

# IV. REGRESSION QUALITATIVE.

## IV.1. Introduction.

On sait qu'il y a deux écoles de pensée en analyse statistique des données.

Pour la première, il y a lieu de se référer à un modèle mathématique de comportement, ayant le plus souvent un caractère linéaire et géométrique, en admettant autour de ce dernier des fluctuations d'échantillonnage.
Ainsi en est-il pour les méthodes de régression linéaire où on cherche à "expliquer" une variable privilégiée retenue d'avance, en fonction d'autres variables dites "explicatives" dont la délimitation constitue un important problème. Les méthodes permettent dans des situations particulières de résoudre le problème.
Dans le cas où les variables sont qualitatives, les méthodes de régression qualitative basée sur le $\Phi^2$ fournissent une solution intéressante au problème posé. Mais, elles supposent une représentation spatiale de variables discrètes qui ne s'impose pas nécessairement et des problèmes de dimension et de choix des prédicteurs inhérents à cette représentation [J. J. DAUDIN (1980)].

Notre propre démarche est conforme à la deuxième école de pensée où on propose une structure de condensation de l'information statistique, assez générale et adaptée à la mise en évidence du phénomène étudié (la régression par exemple) et ce, sans se préoccuper de la pureté de son expression mathématique.
Ce qui est proposé ici et qui résulte directement du travail de B. Tallur [TALLUR (1982a)] correspond surtout à une méthode d'interprétation d'une classification par proximité sur un ensemble d'attributs-modalités de variables qualitatives totalement ordinales dont on distingue une variable "cible" qu'il s'agit de « comprendre » par rapport aux autres

variables. Un des intérêts de la classification hiérarchique des attributs - modalités est de reconnaître directement et automatiquement quelles sont les variables dont le comportement « explique » celui de la variable "cible". Dans le cas de l'exemple concret que nous allons mentionner, la variable à « expliquer » est la Tension Artérielle Systolique (TAS) et les variables « explicatives » peuvent être à caractère biologique ou sociologique.

IV.2 - Exemple concret.

Plusieurs milliers de "bilans de santé" sont réalisés chaque année dans chacun des Centres d'Examens de Santé (C.E.S.) en France, dont les objectifs sont la prévention et le dépistage à temps des maladies. L'Hyper Tension Artérielle (H.T.A.) est considérée comme un des facteurs essentiels du risque cardiovasculaire ; les maladies cardiovasculaires étant la cause du plus grand nombre de décès. D'où la préoccupation considérable des C.E.S. pour la prévention de ces maladies.

Le "bilan de santé" consiste en un questionnaire détaillé rempli par chaque sujet sur ses conditions socio-professionnelles d'une part, en une suite d'examens cliniques et de tests biologiques d'autre part. Les résultats de toutes les analyses sont suivies de la conclusion du médecin qui reconvoquera le sujet en cas d'anomalie quelconque.

Nombreuses études ont montré l'existence de relations entre certaines variables biologiques et celle tensionnelle. Il est dans ces conditions naturel de chercher par une méthode statistique <u>globale</u> les indicateurs de risques cardiovasculaires parmi les facteurs biologiques ou d'ailleurs sociologiques liés à l'H.T.A.. Alors que l'exploitation de ces données précieuses s'est en général limité à l'analyse de l'association des variables deux à deux (coefficient de corrélation, test du $\chi^2$, etc ...)

La présente étude porte sur l'ensemble de quatre C.E.S.: Albi, Nice, Rennes et St-Brieuc. Parmi les consultants de l'année 1979, ont été retenus 10.693 sujets non médicalisés ; en d'autres termes, les sujets soumis à un traitement médical après un dépistage d'anomalie cardiovasculaire ont été éliminés. En effet le but est de définir le seuil critique de la T.A.S. — chez les sujets sains — au delà duquel on peut craindre les risques cardiovasculaires et de dégager des facteurs ou indicateurs de risques.

Le traitement statistique de l'étude a été effectué par B. Tallur et l'interprétation en collaboration avec les médecins M. Caillet, L. Massé, H. Courcous, E. Coste, E. Abou et B. Dupont [Juin 1981].

La population de 10.693 sujets examinés est divisée en quatre sous-populations suivant le sexe et les deux tranches d'âge 30 à 39 ans et 40 à 49 ans. Les effectifs de ces sous-populations sont définis dans le tableau suivant :

| HOMMES | | FEMMES | |
|---|---|---|---|
| 30 à 39 ans | 40 à 49 ans | 30 à 39 ans | 40 à 49 ans |
| 2.924 | 2.531 | 2.855 | 2.383 |

Chaque sous population a été étudiée séparément et les résultats ont été comparés.

Comme nous l'avons mentionné ci-dessus, les variables retenues par les spécialistes sont de deux sortes, sociologiques et biologiques.

a. Variables sociologiques.

Elles sont relatives à la situation socio-professionnelle et au mode de vie familiale du sujet. Les variables descriptives sont toutes qualitatives à l'exception de "quantité de tabac" et "durée de tabagisme" qui ont été découpées cha-

cune en intervalles dont chacun définira une modalité d'une variable qualitative ordinale. Les variables retenues sont les suivantes :

1. situation de famille (7 modalités)
2. catégorie socio-professionnelle (9 modalités)
3. horaire de travail (7 modalités)
4. type d'habitat (9 modalités)
5. mode d'alimentation (4 modalités)
6. catégorie de fumeur (4 modalités)
7. durée de tabagisme (4 modalités)
8. quantité cumulée de tabac (5 modalités)
9. consommation d'alcool (5 modalités).

b. Variables biologiques.

Elles sont toutes quantitatives et continues, toutefois la nécessité d'obtenir des seuils conduit naturellement à la discrétisation de chacune des variables, en découpant de façon adéquate son intervalle de variation en sous intervalles. D'autre part, on se ramène ainsi à des variables qualitatives ordinales de même type que certaines de celles ci-dessus. Les variables biologiques retenues sont les suivantes :

1. Tension Artérielle Systolique (8 modalités)
2. Taux de glycémie en g/l (7 modalités)
3. Taux de cholestérol en g/l (7 modalités)
4. Taux de l'Acide Urique en g/l (7 modalités)
5. Gamma G.T. (7 modalités)
6. Taux des Triglycérides en g/l (8 modalités)
7. Volume Globulaire Moyen (V.G.M.) (7 modalités)
8. Taille (4 modalités)
9. Indice de Quetelet (c'est le rapport du poids en kg sur le carré de la taille en mètres) (8 modalités).

Ce sont des études préliminaires qui ont permis la délimitation des bornes du découpage en classes, relativement à chacune

des variables quantitatives. Toutefois, à partir d'une suggestion que nous lui avions faite, J. Y. Lafaye [LAFAYE (1979a) (1979b)] a mis au point une technique automatique de discrétisation de variables statistiques numériques, valable même pour les « petits » échantillons, qui a été appliquée avec beaucoup d'intérêt à des données médicales [KERJAN (1978)]. Cette technique originale est basée sur la délimitation de zones stables de faible densité de l'intervalle de variation d'un même paramètre.

On aurait pu commencer par une classification hiérarchique des variables qualitatives prises globalement, en utilisant un indice de proximité entre variables définissant chacune une partition sur l'ensemble des sujets [LERMAN (1981a) Chap. 2, VILLOING (1980)] ; en effet, on aura été conduit pour une telle classification à oublier le caractère ordinal des variables qualitatives puisqu'il y a un mélange de variables qualitatives nominales et ordinales.

Toutefois, le problème de définition de seuils trouve plus directement réponse dans la classification hiérarchique d'attributs - modalités correspondant aux différentes variables qualitatives retenues. Chacune des quatre sous-populations précédemment définies donne lieu à un tableau d'incidence qui est un tableau disjonctif - complet, croisant l'ensemble I des sujets avec l'ensemble des 117 modalités des variables.

IV.3 - Principe d'interprétation.

L'Algorithme de la Vraisemblance du Lien (cf. § III.1) a été utilisé sur l'ensemble A des attributs - modalités à partir d'un indice d'association entre attributs (cf. § I.1) conforme à l'h.a.l. $N_3$. Il aurait été certainement intéressant, comme cela a pu être avec succès pratiqué sur d'autres études, d'utiliser l'Algorithme Basé sur la Corrélation (cf. § III.2).

Imaginons qu'on ait retenu - compte tenu du comportement de la statistique globale $\Sigma$ (cf. § III.3) et de l'avis du spécialiste des données - un des niveaux les plus significatifs de l'arbre des classifications, qui définit une partition $\{A_j / 1 \leqslant j \leqslant k\}$ sur A. Chaque $A_j$ définit un profil de comportement qu'on peut plus ou moins aisément exprimer.

Le principe d'une interprétation régressive de la classification, par rapport à une variable "cible", mis en forme par B. Tallur [TALLUR (1982a)], est simple. On commence par regarder comment se répartissent les modalités de la variable à «expliquer», ici la T.A.S., dans les différentes classes $A_j$, $1 \leqslant j \leqslant k$. A cet égard on sait que deux modalités exclusives d'une même variable ne s'associent dans une même classe que par enchaînements avec des modalités d'autres variables (cf. § II.2.2.1.).

On dira qu'on a une <u>correspondance ordinale connexe</u> entre une variable qualitative ordinale <u>a</u> et la typologie $\{A_j / 1 \leqslant j \leqslant k\}$ si le sous-ensemble des modalités de <u>a</u> qui se retrouvent dans une même classe $A_j$, forme un intervalle connexe dans l'ensemble des modalités de <u>a</u> et ce, pour tout $j = 1, 2, \ldots, k$.

Dans le cas où une telle correspondance est mise en défaut, on associera à chaque $A_j$, l'intervalle des modalités de <u>a</u> contenu dans $A_j$ et auquel appartient la modalité la moins "neutre", relativement à une visée classificatoire de l'ensemble des attributs-modalités.

Nous avons en effet introduit un coefficient qu'avant toute classification, on attache à chaque élément de l'ensemble à organiser et qui représente sa variance des proximités au sens de Q' (cf. formule (4) § III.1) aux autres éléments de l'ensemble à classifier [LERMAN (1970), (1981a)] :

$$(\forall j \in J),\quad C(j) = \frac{1}{(\mathrm{card}(J)-1)} \sum_{h \neq j} [\varphi'(j,h) - \varphi'(j)]^2 , \qquad (1)$$

où $$\varphi'(j) = \frac{1}{(\mathrm{card}(J)-1)} \sum_{h \neq j} \varphi'(j,h)$$

La petitesse de $C(j)$ mesure le "degré de neutralité" de $j$. Il a d'ailleurs pu être expérimentalement observé que plus $C(j)$ est grand, plus $j$ intervient de façon significative dans la formation et l'entraînement de la classe où il apparaît.

On commence de la sorte par définir, à partir de la variable $\subseteq$ à «expliquer» (la T.A.S. dans notre cas), un ordre sur l'ensemble des classes $A_j$, ce dernier peut être partiel car il peut exister des classes $A_j$ sans aucune modalité de la variable "cible". Il faut espérer qu'à des modalités relativement "neutres" près, la "correspondance ordinale" est connexe, ce qui est le cas dans notre exemple concret.

Une variable $\underline{a}$ sera appelée "indicateur" de la variable $\subseteq$ si l'ordre total induit (par la procédure précédente) sur l'ensemble des classes $A_j$ deux à deux comparables pour $\subseteq$, est soit identique, soit opposé à celui défini par la variable $\subseteq$. Dans le premier cas $\underline{a}$ est un "indicateur positif" et dans le second, un "indicateur négatif".

Ainsi pour que ce type d'analyse ait un intérêt, il importe que la variable "cible" entraîne dans son sillage un ensemble de variables dont les variations respectives sont concomitantes (positivement ou négativement) et qui seront les plus «explicatives».

Pour les autres variables qui ne sont pas strictement des "indicateurs", mais dont le rôle «explicatif» peut ne pas être négligeable, on peut comme le propose d'ailleurs B. Tallur, calculer la valeur de l'indice d'association présenté au paragraphe

I.2, entre chacune de ces variables et la variable à «expliquer». Nous présenterons au paragraphe suivant IV.5, une mesure globale de l'association entre la variable qualitative ordinale "câble" et la suite ordonnée conformément (cf. ci-dessus) des classes $A_j$ d'attributs.

## IV.4- Application sur l'exemple concret.

Nous reprenons ici les résultats concernant la sous-population de femmes âgées de 30 à 39 ans. Relativement à l'arbre hiérarchique obtenu par A.V.L. sur l'ensemble des attributs modalités, la "statistique globale $E_r$" (cf. § III.3.) atteint son maximum absolu au 89-ème niveau de l'arbre qui définit une partition en quatre grandes classes notées $A_1$, $A_2$, $A_3$ et $A_4$. On constate que (voir figure de l'arbre condensé des classifications) :

- $A_1$ ne contient aucune modalité de T.A.S.
- $A_2$ contient "TAS<10" et "TAS 10 à 12,9" ; on lui associe l'intervalle "TAS≤12,9".
- $A_3$ contient la modalité "TAS 14 à 14,9" qui lui est associée
- $A_4$ contient les modalités "TAS 13 à 13,9", "TAS 15 à 15,9", "TAS 16 à 16,9", "TAS 17 à 17,9" et "TAS≥18". Il s'agit d'associer soit la modalité "TAS 13 à 13,9", soit l'intervalle "TAS≥15". Selon le critère du degré de neutralité, c'est l'intervalle "TAS≥15" qui est associé à $A_4$.

L'ordre défini sur l'ensemble des classes par la TAS est le suivant :

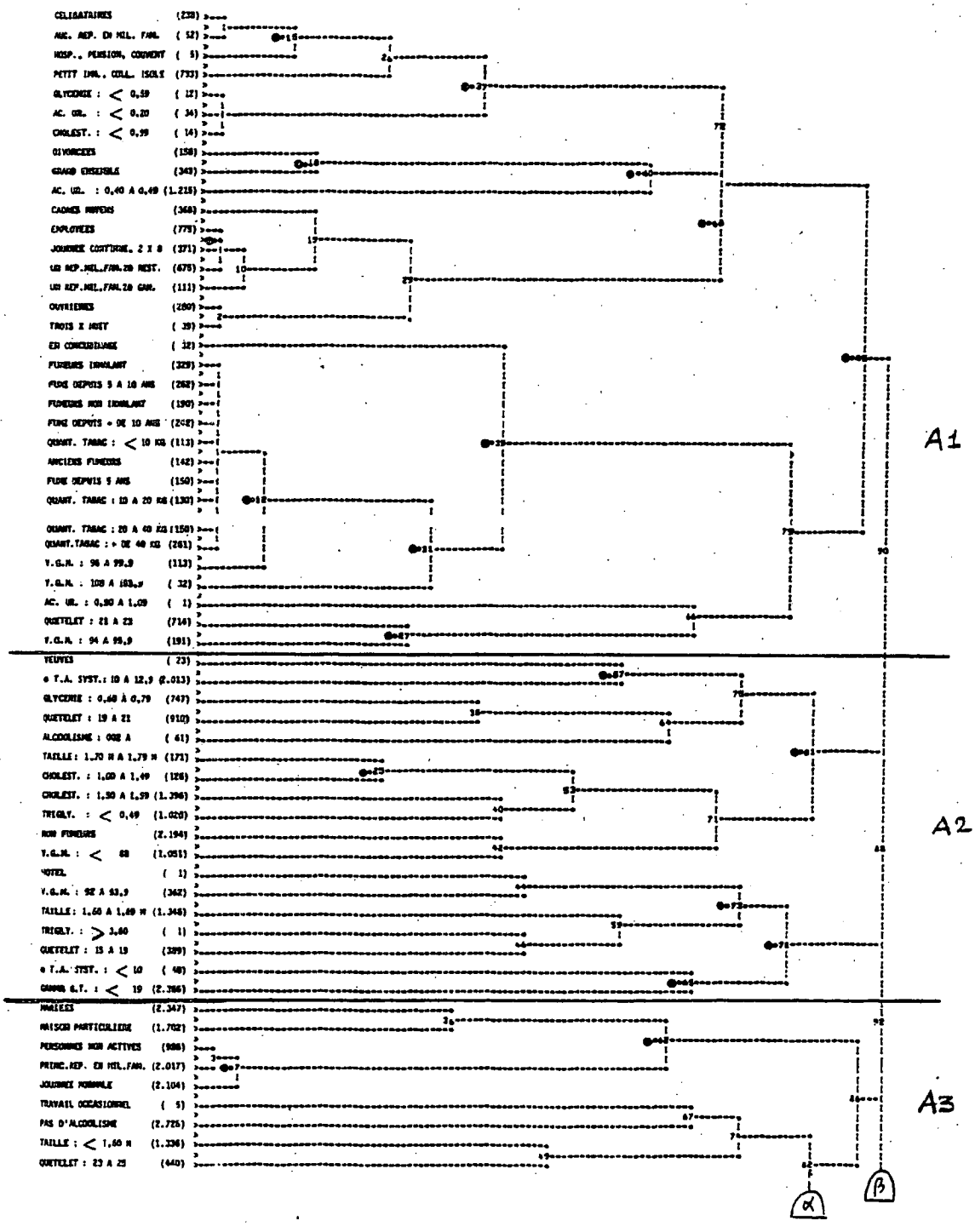
$A_2 < A_3 < A_4$

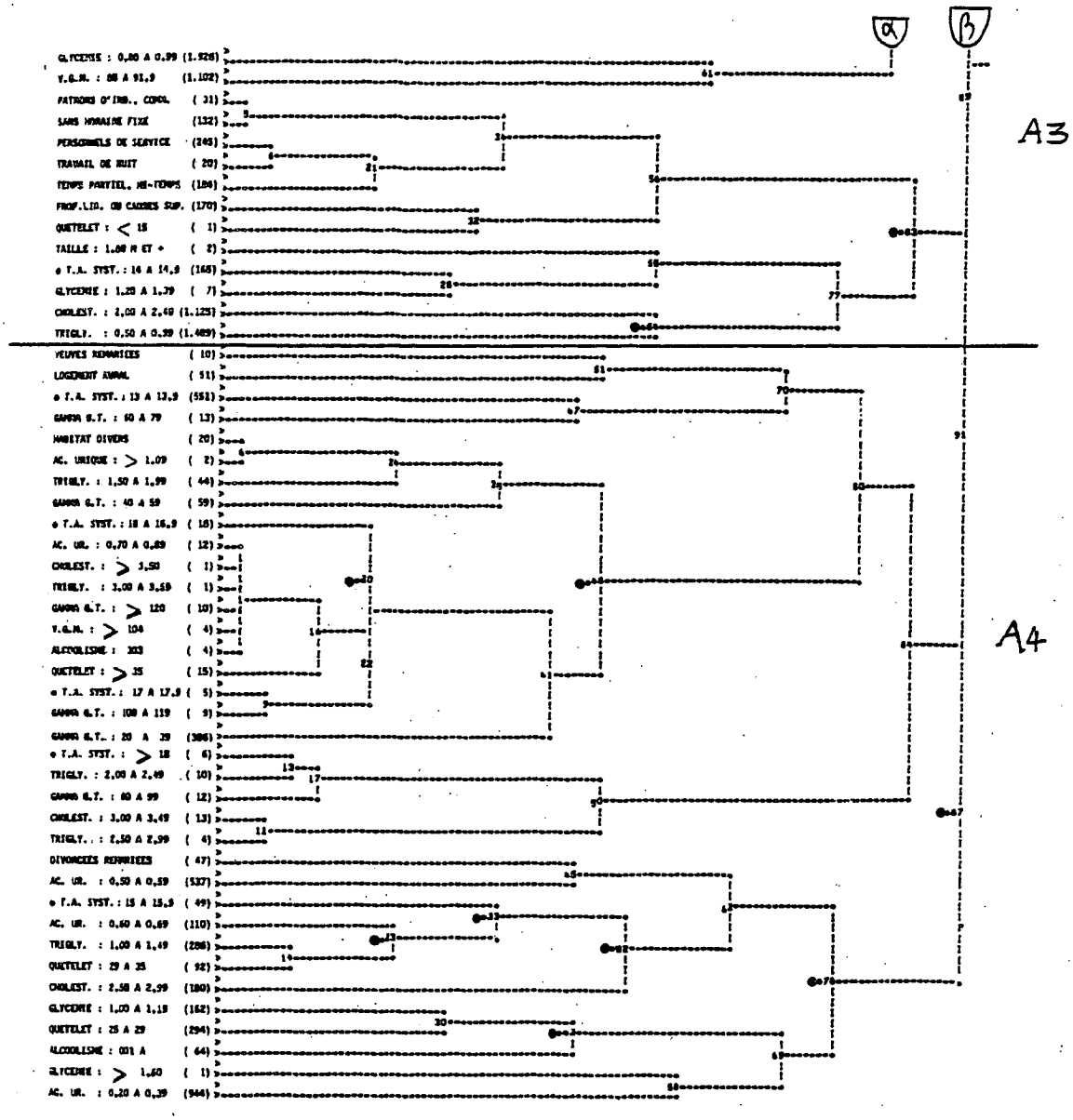
$A_1$ n'étant pas comparable aux autres classes. Le tableau 1 résume le résultat de ces démarches pour quelques unes des variables ordinales les plus liées à la TAS.

On constate que Taux de Cholestérol, Taux de Triglycérides, Taux de Glycémie et Indice de Quetelet sont des "indicateurs positifs" de la T.A.S. ; en effet chacune de ces variables ordonne les classes $A_2$, $A_3$ et $A_4$ de la même manière ($A_2 < A_3 < A_4$). Par ailleurs, bien que Gamma G.T. n'est pas au sens strict du terme, un indicateur ; l'ordre partiel

REPRÉSENTATION DE L'ARBRE

— HYPERTENSION ARTÉRIELLE — 4 CENTRES { ALBI, NICE, SAINT-BRIEUC, RENNES } — FEMMES DE 30 A 39 ANS —






(Les nombres entre parenthèses sont des effectifs)

Figure 5

induit sur l'ensemble des classes $\{A_j / 1 \leq j \leq 4\}$ reste compatible avec celui défini par la T.A.S. puisqu'on a $A_2 < A_4$ pour la Gamma G.T.

a) Mélange des variables explicatives, ordinales et non-ordinales

La présence des variables à l'ensemble des modalités non-ordonné n'affecte pas la méthode de la recherche des indicateurs ; on peut même envisager d'ordonner l'ensemble des modalités d'une telle variable, en affectant à chaque modalité le rang de la classe -défini par la TAS- qui la contient. Ainsi, pour la variable "situation de famille", les modalités sont réparties sur l'ensemble des classes de la façon suivante :

$A_1$ : célibataire, divorcée, en concubinage
$A_2$ : veuve
$A_3$ : mariée
$A_4$ : veuve remariée, divorcée remariée

Etant donné que $A_2 < A_3 < A_4$ pour la T.A.S., on peut ordonner les modalités de "situation de famille" par rapport à la T.A.S. comme suit :

Veuve<Mariée <(Veuve remariée = Divorcée remariée)

| Variables | T.A.S. | | | Taux de cholestérol | | | Triglycérides | | | Glycémie | | | Indice de Quetelet | | | Gamma G.T. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classes | Modalité associée | Rang | Restriction du rang | Modalité associée | Rang | Restriction du rang | Modalité associée | Rang | Restriction du rang | Modalité associée | Rang | Restriction du rang | Modalité associée | Rang | Restriction du rang | Modalité associée | Rang | Restriction du rang |
| $A_1$ | aucune | - | | ≤0,99 | 1 | | aucune | - | | ≤0,59 | 1 | | 21-23 | 2 | | aucune | - | |
| $A_2$ | ≤12,9 | 1 | | 1-1,99 | 2 | 1 | ≤0,49 | 1 | 1 | 0,60-0,79 | 2 | 1 | 15-21 | 1 | 1 | ≤19 | 1 | 1 |
| $A_3$ | 14-14,9 | 2 | | 2-2,49 | 3 | 2 | 0,50-0,99 | 2 | 2 | 0,80 à 0,99 | 3 | 2 | 23-25 | 3 | 2 | aucune | - | - |
| $A_4$ | ≥15 | 3 | | ≥2,50 | 4 | 3 | 1-3,59 | 3 | 3 | 1,00 à 1,19 | 4 | 3 | ≥25 | 4 | 3 | ≥20 | 2 | 2 |

TABLEAU 1 : Résultats des démarches conduisant à la découverte des indicateurs de la T.A.S.

La classe $A_1$ se joignant à $A_2$ au niveau 90 de l'arbre, on pourra éventuellement définir trois classes des modalités ordonnées :

(Célibataire=divorcée=concubinage=veuve)<Mariée

<(Veuve remariée = divorcée remariée)

b) Interprétation des classes

La classe $A_1$ regroupe toutes les modalités concernant les fumeuses, telles que "fume depuis 5 ans", "inhale la fumée", "quantité de tabac<10 kg",...etc. Cette classe ne contient pas de modalités de TAS. Il y a par contre quelques faibles valeurs des variables biologiques telles que cholestérol et glycémie. Il semble donc que le fait de fumer n'influence pas sur la TAS, et que certain nombre de paramètres biologiques sont faibles chez les fumeuses. L'interprétation dynamique de l'arbre aux niveaux supérieurs montre que (voir figure 1) cette classe se réunit avec la classe $A_2$ qui est caractérisée par les valeurs moyennes de la TAS ainsi que celles de tous les paramètres biologiques. Il s'agit donc d'une classe "normale". Remarquons que $A_2$ contient "non-fumeuses".

$A_3$ et $A_4$, qui se réunissent d'ailleurs au niveau 91, sont composées respectivement des valeurs élevées et très élevées de toutes les variables biologiques ainsi que celle de la T.A.S. La partition au niveau 91 en deux classes sépare toutes les valeurs de TAS inférieures à 13 de celles qui sont ≥13 ; ceci montre qu'il existe un seuil entre les valeurs faibles ou "normales" et les valeurs fortes de la TAS et que ce seuil se situe autour de 13. Pour la médecine préventive, la connaissance de ce seuil est très important, car elle permet de mieux surveiller les sujets au delà de cette valeur accompagnés d'autres symptômes indicateurs de risques.

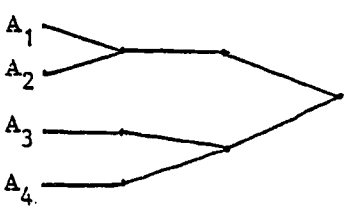



Figure. 6

Le même seuil a été retrouvé par B. Tallur à partir d'une classification hiérarchique par l'A.V.L. de l'ensemble des huit modalités de la T.A.S. à travers l'ensemble des 109 modalités des autres variables [TALLUR (1982a)]. Il s'agit ici d'une structure de la donnée qui correspond à une juxtaposition horizontale de tables de contingences (cf. § II.1, expression (2)). L'indice de départ de comparaison entre deux modalités $i$ et $i'$ de la T.A.S. correspond à une adaptation de l'indice $\rho(i, i')$ (cf. formule (11) § II.2.1., fin § II.2.3. et LERMAN, TALLUR (1980))

Signalons que cette dernière approche a, relativement à la classification hiérarchique, des affinités avec ce que propose P. Cazes dans une méthode dite de régression "par boules", à partir de l'analyse factorielle des correspondances [CAZES (1976)]

## IV.5 - Qualité de la régression

Pour mesurer la qualité de l'adéquation ordinale entre la suite des modalités de la variable à « expliquer » et les classe $A_j$, $1 \leq j \leq k$, d'attributs où elles apparaissent, nous sommes conduits à procéder de la manière suivante.

Pour tout $j$ où au moins une modalité de la variable "cible" apparaît dans la classe $A_j$, faire :

a) substituer à chaque classe $A_j$ une nouvelle classe $\tilde{A}_j$ obtenue

- en extrayant de $A_j$ l'ensemble des attributs-modalités - de chaque variable qualitative ordinale - qui ne sont pas conformes à la condition de la "correspondance ordinale connexe" (cf. § IV.3 ci-dessus) et en réunissant les autres - qui forment un intervalle incluant la modalité la moins "neutre" - en un seul « gros » attribut.

- regrouper en un seul « gros » attribut les modalités d'une même variable qualitative nominale, qui ap-

paraissent dans $A_j$ .

b) ôter de la nouvelle classe $\tilde{A}_j$ , le nouvel attribut $\tilde{c}_j$ relatif à l'intervalle des valeurs de la variable "cible" attribuée à la classe $A_j$ ; la classe résultante sera notée $B_j = \tilde{A}_j - \{\tilde{c}_j\}$.

Il s'agit dans ces conditions de mettre en correspondance la suite ordonnée des modalités $\tilde{c}_j$ et celle associée des classes $B_j$ d'attributs-modalités, puis de proposer une mesure de l'intensité de leur association dont la valeur définira la qualité de la régression.

Si on n'avait pas à tenir compte de l'ordre, notre problème se trouve résolu au moyen d'un indice d'association que nous avons mis au point et développé entre la classification "nette" définie par l'ensemble des modalités $\tilde{c}_j$ et celle "floue" - définie - au moyen d'un "degré d'appartenance" (cf. §II.2.3) - par l'ensemble des classes $B_j$. [LERMAN (1979), (1981a) Chap. 3] où nous étudions de plus le cas le plus général du croisement de deux classifications "floues".

La situation qui se présente à nous ici de comparaison d'une variable qualitative ordinale "nette" avec une variable qualitative ordinale "floue" - dont chaque modalité se trouve définie par une classe d'attributs - est donc nouvelle et originale. Nous allons dans ces conditions présenter la nature des calculs qui ont été effectués [LERMAN (1983c)] pour obtenir l'indice d'association. A cette fin, il est nécessaire de reprendre clairement nos notations compte tenu du fait que certains indices $j$, pour $j = 1, 2, \ldots, k$, peuvent ne pas être atteints si $A_j$ ne contient aucune modalité de la variable à « expliquer ».

$I = \{1, 2, \ldots, i, \ldots, n\}$ désignera ici l'ensemble des in-

dices des sujets qui ont répondu aux différentes modalités ou attributs qui interviendront ci-dessous.

$L = \{1, 2, \ldots, \ell, \ldots, M(\ell)\}$ est l'ensemble des indices des modalités de la variable qualitative ordinale "nette" qu'on notera c et dont une modalité courante est ainsi désignée par $c_\ell$ , $1 \leqslant \ell \leqslant M(\ell)$.

$\Lambda = \{1, 2, \ldots, \lambda, \ldots, M(\lambda)\}$ est l'ensemble des indices des modalités de la variable qualitative ordinale "floue" qu'on peut noter $\beta$ et dont une modalité ("floue") courante est ainsi désignée par $\beta_\lambda$ , $1 \leqslant \lambda \leqslant M(\lambda)$. Rappelons que $\beta_\lambda$ se trouve en fait définie par une classe $B_\lambda$ d'attributs "orientés" (i.e. telle que deux attributs correspondants à deux modalités exclusives ne peuvent tous les deux y appartenir).

La démarche que nous allons emprunter est toujours la même (cf. § I). Ce qui va changer par rapport à la comparaison de deux variables qualitatives ordinales "nettes" (se reporter au paragraphe I.2) est le remplacement des fonctions indicatrices $\varphi_{D(m)}$ , $1 \leqslant m \leqslant M$ , par des fonctions d'appartenance $\varphi_{B_\lambda}$ (cf. formule (5) § II.1), $1 \leqslant \lambda \leqslant M(\lambda)$. Toutefois les calculs, surtout celui de la variance de la v.a. associée à l'indice brut, devient sensiblement plus délicat.

L'indice brut d'association entre les deux variables c et $\beta$ se met dans ces conditions sous la forme

$$s(c,\beta) = \sum \{ \sum \{ [\varphi_\ell(i)\varphi_{\ell'}(j)][\varphi_\lambda(i)\varphi_{\lambda'}(j)] \,/\, (i,j) \in I^{[2]} \} \,/ \quad (2)$$
$$1 \leqslant \ell < \ell' \leqslant M(\ell),\ 1 \leqslant \lambda < \lambda' \leqslant M(\lambda) \},$$

où $I^{[2]}$ désigne l'ensemble des couples à composantes distinctes de I.

Nous considérons une forme unilatérale de l'h.a.l. où, au préordre total sur I défini par la variable c et dont la suite des classes peut être notée $\{C(\ell) \,/\, 1 \leqslant \ell \leqslant M(\ell)\}$, on associe

dans l'ensemble des préordres totaux de même composition et muni d'une probabilité uniforme, un préordre total aléatoire $C'$ dont la suite des classes peut être notée $\{C'(\ell)/1\leqslant\ell\leqslant M(\ell)\}$. Ainsi, on substituera dans l'h.a.l. à la fonction indicatrice (f.i.) $\varphi_\ell$ de $C_\ell$, la f.i. aléatoire $\varphi'_\ell$ de $C'(\ell)$.

Il importe certes de définir une h.a.l. duale et d'y effectuer le même type de calculs

L'espérance mathématique de la v.a. $S(C',\beta)$ s'obtient sans difficulté, elle est égale à

$$\text{moy.}[S(C',\beta)]=\frac{1}{n(n-1)}\sum\{n(\ell)n(\ell')[\nu(\lambda)\nu(\lambda')-\nu_{11}(\lambda,\lambda')]/ \quad (3)$$

$$1\leqslant\ell<\ell'\leqslant M(\ell),\ 1\leqslant\lambda<\lambda'\leqslant M(\lambda)\},$$

où $\nu(\lambda)=\sum\{\varphi_\lambda(i)/i\in I\}$ et $\nu_{11}(\lambda,\lambda')=\sum\{\varphi_\lambda(i)\varphi_{\lambda'}(i)/i\in I\}$.

Le calcul de la variance suppose celui du moment absolu d'ordre 2 qui s'obtient à partir d'une décomposition spécifique du carré de la v.a. $S(C',\beta)$ :

L'ensemble d'indexation de la somme définissant $S^2(C',\beta)$ se trouve défini par le carré cartésien de

$$I^{[2]}\times L^{\{2,4\}}\times\Lambda^{\{2,4\}} \quad (4)$$

où nous avons noté $L^{\{2,4\}}$ (resp. $\Lambda^{\{2,4\}}$) l'ensemble des paires d'indices $\{\ell,\ell'\}$ (resp. $\{\lambda,\lambda'\}$) où $\ell<\ell'$ (resp. $\lambda<\lambda'$).

La somme définissant $S^2(C',\beta)$ est ainsi indexée par

$$(I^{[2]}\times I^{[2]})\times(L^{\{2,4\}}\times L^{\{2,4\}})\times(\Lambda^{\{2,4\}}\times\Lambda^{\{2,4\}}), \quad (5)$$

sa décomposition s'effectue selon le croisement de trois partitions respectivement définies sur

$$I^{[2]}\times I^{[2]},\quad L^{\{2,4\}}\times L^{\{2,4\}}\text{ et }\Lambda^{\{2,4\}}\times\Lambda^{\{2,4\}},$$

où chacune des partitions est définie selon la structure du

4-uplet considéré. Plus précisément :

La partition de $I^{[2]} \times I^{[2]}$ se fait selon les ensembles

$D(I) = \{((i,j),(i,j))\}$ de cardinal $n(n-1)$

$J(I) = \{((i,j),(j,i))\}$ " " $n(n-1)$

$G_1(I) = \{((i,j),(i,k))\}$ " " $n(n-1)(n-2)$

$G_1'(I) = \{((i,j),(k,i))\}$ " " $n(n-1)(n-2)$

$G_2(I) = \{((i,j),(k,j))\}$ " " $n(n-1)(n-2)$

$G_2'(I) = \{((i,j),(j,k))\}$ " " $n(n-1)(n-2)$

$H(I) = \{((i,j),(k,l))\}$ " " $n(n-1)(n-2)(n-3)$,

où des lettres différentes indiquent des indices différents.

La partition de $L^{124} \times L^{124}$ se fait selon les ensembles

$\tilde{D}(L) = \{((p,q),(p,q))\}$ – où $p<q$ – de cardinal $m(m-1)/2$

$\tilde{G}_1(L) = \{((p,q),(p,s))\}$ – où $p<q$ et $p<s$ – " " $(m-2)(m-1)m/3$

$\tilde{G}_1'(L) = \{((p,q),(r,p))\}$ – où $p<q$ et $r<p$ – " " $(m-2)(m-1)m/6$

$\tilde{G}_2(L) = \{((p,q),(r,q))\}$ – où $p<q$ et $r<q$ – " " $(m-2)(m-1)m/3$

$\tilde{G}_2'(L) = \{((p,q),(q,s))\}$ – où $p<q$ et $q<s$ – " " $(m-2)(m-1)m/6$

$\tilde{H}(L) = \{((p,q),(r,s))\}$ – où $p<q$ et $r<s$ – " " $(m-3)(m-2)(m-1)m/4$

où des lettres différentes indiquent des indices différents et où on a noté $m = M(l)$.

La partition de $\Lambda^{124} \times \Lambda^{124}$ est analogue à la précédente; ses classes sont naturellement notées

$\tilde{D}(\Lambda)$, $\tilde{G}_1(\Lambda)$, $\tilde{G}'_1(\Lambda)$, $\tilde{G}_2(\Lambda)$, $\tilde{G}'_2(\Lambda)$ et $\tilde{H}(\Lambda)$.

La première décomposition qu'on considère est conforme à la ci-dessus partition de $I^{[2]} \times I^{[2]}$.

On peut déjà remarquer que la classe $J(I)$ donne lieu à une contribution nulle ; en effet $\varphi_\ell(i)\varphi_{\ell'}(j)\varphi_p(j)\varphi_{p'}(i)$ est nécessairement nul pour $\ell<\ell'$ et $p<p'$.

Nous allons, pour dévoiler la nature des calculs, indiquer la contribution de la classe $G_1(I)$. Cette dernière peut se mettre sous la forme de l'espérance de

$$\sum_{G_1(I)} \Big\{ \Big[ \Big( \sum_{\ell<\ell'} \varphi^*_\ell(i)\varphi^*_{\ell'}(j) \Big) \Big( \sum_{p<p'} \varphi^*_p(i)\varphi^*_{p'}(k) \Big) \Big] \Big[ \Big( \sum_{\lambda<\lambda'} \varphi_\lambda(i)\varphi_{\lambda'}(j) \Big) \Big( \sum_{\mu<\mu'} \varphi_\mu(i)\varphi_{\mu'}(k) \Big) \Big] \Big. \tag{6}$$

L'espérance mathématique du contenu du premier crochet a été déterminée dans le cas de la comparaison de deux variables qualitatives ordinales nettes [LERMAN (1973), (1981a) et (1983c)]. Il reste à évaluer

$$\sum_{G_1(I)} \Big[ \Big( \sum_{\lambda<\lambda'} \varphi_\lambda(i)\varphi_{\lambda'}(j) \Big) \Big( \sum_{\mu<\mu'} \varphi_\mu(i)\varphi_{\mu'}(k) \Big) \Big] \tag{7}$$

et c'est là qu'interviènt la décomposition conformément à la ci-dessus partition de $\Lambda^{[2]} \times \Lambda^{[2]}$.

Ainsi, la partition de $L^{[2]} \times L^{[2]}$ n'intervient pas directement et on aura à préciser 36 expressions conformément au croisement des deux partitions sur $I^{[2]} \times I^{[2]}$ et sur $\Lambda^{123} \times \Lambda^{123}$. Le détail des calculs paraîtra prochainement dans [LERMAN (1983c)] ; d'autre part, la conception d'un programme par Mlle Moreau (Stagiaire 3ème cycle est en cours).

# BIBLIOGRAPHIE.

J.P. BENZECRI ; "Sur les algorithmes de classification", Cours ISUP. Paris, 1965_1966.

J.P. BENZECRI ; "L'analyse des données" (tomes 1 et 2), (1973), dernière édition (1980), Dunod, Paris.

J.L. BUARD ; "Gestion statistique des demandes d'actes biologiques. Typologie des unités fonctionnelles". Thèse de 3ème cycle (Traitement de l'Information), Université de Rennes I, (1980).

M. CAILLET, L. MASSE, H. COURCOUX, E. COSTE, E. ABOU, B. DUPONT et B. TALLUR ; "Importance du niveau de tension artérielle systolique dans la selection de populations cibles en médecine préventive.", Congrès d'Epidémiologie de Bordeaux, Juin (1981).

J.J. DAUDIN ; "Régression qualitative : choix de l'espace prédicteur", in "Data Analysis and Informatics" (E. Diday, L. Lebart, J.P. Pagès and R. Tomassone : editors), North-Holland (1980).

A.M. KERJAN ; "Tentative d'établissement de cent typologies d'examens biologiques. Contribution à l'établissement du système "A.D.M." ", Thèse de doctorat de médecine, Université de Rennes, (1978).

J.Y. LAFAYE ; "Les différentes formes de l'appréhension des données dans l'exploration fonctionnelle hépatique ; discrétisation de variables numériques. Recherche de profils biologiques par une méthode de classification hiérarchique", Thèse de 3ème cycle (Traitement de l'Information), Université de Rennes I, (1978).

J.Y. LAFAYE ; "Une méthode de discrétisation d'une variable continue" Rev. Stat. Appl. n°2 (1979a).

J.Y. LAFAYE; "Une méthode automatique de discrétisation de variables numériques représentées par de petits échantillons", Actes du Congrès AFCET : « Reconnaissance des formes et intelligence artificielle », Toulouse Sept. (1979 b).

L. LEBART, A. MORINEAU et N. TABARD; "Techniques de la description statistique", (1977), (dernière édition (1982)), Dunod, Paris.

I.C. LERMAN; "Sur l'analyse des données préalable à une classification automatique; proposition d'une nouvelle mesure de similarité". Rev. Math. & Sc. Hum., 8ème année, n°32, (1970).

I.G. LERMAN; "Etude distributionnelle de statistiques de proximité entre structures finies de même type; application à la classification automatique." Cahiers du B.U.R.O. n° 19, Paris, (1973).

I.G. LERMAN; "Formal analysis of a general notion of proximity between variables" in Proceed. published by North Holland in (1977) of "Congrès Européen des Statisticiens", Grenoble (1976).

I.G. LERMAN; "Croisement de classifications «floues»", Publ. Inst. Stat. Univ. Paris, XXIV, fasc. 1-2, Paris (1979).

I.G. LERMAN; "Classification et analyse ordinale des données", Dunod, Paris (1981a).

I.C. LERMAN; "Sur la signification des classes issues d'une classification automatique de données", à paraître dans les actes de « Advanced Studies Institute on Numerical Taxonomy, Bad Windsheim, Juillet (1982a).

I.C. LERMAN; "Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence", rapport IRISA n° 182, Rennes, (1982).

I.C. LERMAN ; "Indices d'association partielle entre variables qualitatives nominales", - rapport IRISA n°153, Oct. (1981) Rennes - à paraître dans la R.A.I.R.O. série verte (1983 a)

I.C. LERMAN, "Indices d'association partielle entre variables qualitatives ordinales", - rapport IRISA n° 153, Oct. (1981) Rennes - à paraître dans les Publications de l'Institut de Stat. de l'Univ. de Paris (1983 b)

I.C. LERMAN. "Indice d'association entre variables qualitatives ordinales floues", rapport IRISA en cours de fabrication, Rennes (1983 c).

I.C. LERMAN, B. TALLUR ; "Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence", Rev. de Stat. Appl. n° 28, 3, Paris (1980).

A. PROD'HOMME ; "Indice d'explication des classes obtenues par une méthode de classification hiérarchique respectant la contrainte de contiguïté spaciale. Application à la viticulture Girondine et à la construction de logements dans les Bouches du Rhône." Thèse de 3ème cycle, Univ. de Rennes I, Déc. (1980).

B. TALLUR ; "Etude de l'Agriculture régionale Française", rapport IRISA n° 103, Université de Rennes I, (1978).

B. TALLUR ; "Méthode d'interprétation d'une classification hiérarchique d'attributs-modalités pour l'explication d'une variable ; application à la recherche d'un seuil critique de la tension systolique et des indicateurs de risques cardiovasculaires", rapport IRISA n° 159, Université de Rennes I, (1982), (à paraître dans Rev. Stat. Appl.)

B. TALLUR ; "Un nouvel algorithme de classification hiérarchique des éléments constitutifs de tableaux de contingence, basé sur la corrélation", rapport IRISA n° 177, Rennes (1982).

P. VILLOING ; "Classification ascendante hiérarchique et indices de similarité sur données qualitatives nominales selon l'algorithme de la vraisemblance du lien", Thèse de 3ème cycle, Université de Rennes I, Déc. (1980).

# Liste des Publications Internes IRISA

PI 155 **Analyses d'opinions d'instituteurs à l'égard de l'appropriation des nombres naturels par les élèves de cycle préparatoire**
R. Gras , 37 pages ; *Octobre 1981*

PI 156 **Récursion induction principle revisited**
G. Boudol, L. Kott , 49 pages ; *Décembre 1981*

PI 157 **Loi d'une variable aléatoire à valeur $R^+$ réalisant le minimum des moments d'ordre supérieur à deux lorsque les deux premiers sont fixés**
M.Kowalowka, R. Marie , 8 pages : *Décembre 1981*

PI 158 **Réalisations stochastiques de signaux non stationnaires, et identification sur un seul échantillon**
A. Benveniste J.J. Fuchs , 33 pages ; *Mars 1982*

PI 159 **Méthode d'interprétation d'une classification hiérarchique d'attributs-modalites pour l'«explication» d'une variable ; application à la recherche de seuil critique de la tension artérielle systolique et des indicateurs de risque cardiovasculaire**
B. Tallur , 34 pages ; *Janvier 1982*

PI 160 **Probabilite stationnaire d'un réseau de files d'attente multiclasse à serveur central et à routages dépendant de l'état**
L.M. Le Ny , 18 pages ; *Janvier 1982*

PI 161 **Détection séqentielle de changements brusques des caractéristiques spectrales d'un signal numérique**
M. Basseville, A. Benveniste , pages ; *Mars 1982*

PI 162 **Actes regroupés des journées de Classification de Toulouse (Mai 1980), et de Nancy (Juin 1981)**
I.C. Lerman , 304 pages ;

PI 163 **Modélisation et Identification des caractéristiques d'une structure vibratoire : un problème de réalisation stochastique d'un grand système non stationnaire**
M. Prévosto, A. Benveniste, B. Barnouin , 46 pages ; *Mars 182*

PI 164 **An enlarged definition and complete axiomatization of observational congruence of finite processes**
Ph. Darondeau , 45 pages ; *Avril 1982*

PI 165 **Accès vidéotex à une banque de données médicales**
A. Chauffaut, M. Dragone, R. Rivoire, J.M. Roger , 25 pages ; *Mai 1982*

PI 166 **Comparaison de groupes de variables définies sur le même ensemble d'individus**
B. Escofier, J. Pages , 115 pages ; *Mai 1982*

PI 167 **Transport en circuits virtuels internes sur réseau local et connexion Transpac**
M. Tournois, R. Trépos , 90 pages ; *Mai 1982*

PI 168 **Impact de l'intégration sur le traitement automatique de la parole**
P. Quinton , 14 pages ; *Mai 1982*

PI 169 **A systolic algorithm for connected word recognition**
J.P. Banâtre, P. Frison, P. Quinton , 13 pages ; *Mai 1982*

PI 170 **A network for the detection of words in continuous speech**
J.P. Banâtre, P. Frison, P. Quinton , 24 pages ; *Mai 1982*

PI 171 **Le langage ADA : Etude bibliographique**
J. André, Y. Jégou, M. Raynal , 12 pages ; *Juin 1982*

PI 172 **Comparaison de groupes de variables : 2ème partie : un exemple d'application**
B. Escofier, J. Pajès , 37 pages ; *Juillet 1982*

PI 173 **Unfold-fold program transformations**
L. Kott , 29pages ; *Juillet 1982*

PI 174 **Remarques sur les langages de parenthèses**
J.M. Autebert, J. Beauquier, L. Boasson, G. Senizergues , 20 pages ; *Juillet 1982*

PI 175 **Langages de parenthèses, langages N.T.S. et homomorphismes inverses**
J.M. Autebert, L. Boasson, G. Senizergues , 26 pages ; *Juillet 1982*

PI 176 **Tris pour machines synchrones ou Baudet Stevenson revisited**
R. Rannou , 26 pages ; *Juillet 1982*

PI 177 **Un nouvel algorithme de classification hiérarchique des éléments constitutifs de tableau de contingence basé sur la corrélation**
B. Tallur , Juillet 1982 ;

PI 178 **Programmes d'analyse des re'sultats d'une classification automatique**
I.C. Lerman et collaborateurs , 79 pages ; *Septembre 1982*

PI 179 **Attitude à l'égard des mathématiques des élèves de sixième**
J.Degouys, R. Gras, M. Postic , 29 pages ; *Septembre 1982*

PI 180 **Traitements de textes et manipulations de documents : bibliographie analytique**
J. André , 20 pages ; *Septembre 1982*

PI 181 **Algorithme assurant l'insertion dynamique d'un processeur autour d'un réseau à diffusion et garantissant la cohérence d'un système de numérotation des paquets global et réparti**
Annick Le Coz, Hervé Le Goff, Michel Ollivier , 31 pages ; *Octobre 1982*

PI 182 **Interprétation non linéaire d'un coefficient d'association entre modalités d'une juxtaposition de tables de contingence**
Israël César Lerman , 34 pages ; *Novembre 1982*

PI 183 **L'IRISA vu à travers les stages effectués par ses étudiants de DEA (1è année de thèse)**
Daniel Herman , 41 pages ; *Novembre 1982*

PI 184 **Commande non linéaire robuste des robots manipulateurs**
Claude Samson , 52 pages ; *Janvier 1983*

PI 185 **Dialogue et représentation des informations dans un système de messagerie intelligent**
Philippe Besnard, René Quiniou, Patrice Quinton, Patrick Saint-Dizier, Jacques Siroux, Laurent Trilling , 45 pages ; *Janvier 1983*

PI 186 **Analyse classificatoire d'une correspondance multiple ; typologie et régression**
I.C. Lerman , 54 pages ; *Janvier 1983*